Structure-Based Function Prediction of Functionally Unannotated Structures in the PDB
Prediction of ATP, GTP, Sialic Acid, Retinoic Acid and Heme-bound and Unbound (Free) Nitric Oxide Protein Binding Sites

Vicente M. Reyes, Ph.D.*
E-mail: vmrshi.rit.biology@gmail.com

*work done at:
Dept. of Pharmacology, School of Medicine, University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093-0636
&
Dept. of Biological Sciences, School of Life Sciences Rochester Institute of Technology
One Lomb Memorial Drive, Rochester, NY 14623

Abstract: Due to increased activity in high-throughput structural genomics efforts around the globe, there has been a steady accumulation of experimentally solved protein 3D structures lacking functional annotation, thus creating a need for structure-based protein function assignment methods. Computational prediction of ligand binding sites (LBS) is a well-established protein function assignment method. Here we apply the specific ligand binding site detection algorithm we recently described (Reyes, V.M. & Sheth, V.N., 2011; Reyes, V.M., 2015a) to some 801 functionally unannotated experimental structures in the Protein Data Bank by screening for the binding sites of six biologically important ligands, namely: GTP in small Ras-type G-proteins, ATP in ser/thr protein kinases, sialic acid, retinoic acid, and heme-bound and unbound (free) nitric oxide. Validation of the algorithm for the GTP- and ATP-binding sites has been previously described in detail (ibid.); here, validation for the binding sites of the four other ligands shows both good specificity and sensitivity. Of the 801 structures screened, eight tested positive for GTP binding, 61 for ATP binding, 35 for sialic acid binding, 132 for retinoic acid binding, 33 for heme-bound nitric oxide binding, and 10 for free nitric oxide binding. Using the ‘cutting plane’ and ‘tangent sphere’ methods we described previously, (Reyes, V.M., 2015b), we also determined the depth of burial of the ligand binding sites detected above and compared the values with those from the respective training structures, and the degree of similarity between the two values taken as a further validation of the predicted LBSs. Applying this criterion, we were able to narrow down the predicted GTP-binding proteins to two, the ATP-binding proteins to 13, the sialic acid-binding proteins to two, the retinoic acid-binding proteins to 14, the heme-bound NO-binding proteins to four, and the unbound NO-binding proteins to one. We believe this further criterion increases the confidence level of our LBS predictions. The next logical step would be the experimental determination of the actual binding of these putative proteins to their respective ligands.

Keywords: GTP binding site/proteins, ATP binding site/proteins, sialic acid binding site/proteins, retinoic acid binding site/proteins, heme-NO binding site/proteins, unbound NO binding site/proteins, protein function prediction, protein function annotation, protein-ligand interaction(s)
**Abbreviations:** BS, binding site; LBS, ligand binding site; PDB, Protein Data Bank; GTP, guanosine triphosphate; ATP, adenosine triphosphate; SRGP, small Ras-type G-proteins; STPK, ser/thr protein kinase; SIA, sialic acid; REA, retinoic acid; NO, nitric oxide; hNO, heme-bound NO; fNO, free or unbound NO; PLI, protein ligand interaction(s); 3D SM, 3-dimensional search motif; CP, cutting plane; CPM, CP method; TS, tangent sphere; TSM, TS method; CPI, cutting plane index; TSI, tangent sphere index; H-bond, hydrogen bond; VDW, van der Waals; AAR, all-atom representation; DCRR, double-centroid reduced representation; $Z(s)$, the side-chain centroid of amino acid $Z$; $X(b)$, the backbone centroid of amino acid $X$; DCRR, double-centroid reduced representation

1 Introduction.

Progress in both the genomic sequencing efforts around the globe (Burley, S.K. 2000; Heinemann, U. 2000; Terwilliger, T.C., 2000; Norvell, J.C., & Machalek, A.Z., 2000) as well as that of the various high-throughput 3D structure-determination methods (experimental or predicted) of proteins have brought about the accumulation of protein structures which completely lack functional information (Bentley et al., 2004; Murphy et al., 2004; Baxevanis, 2003; Miller et al., 2003). For instance, the Protein Data Bank (PDB), the world’s repository for protein 3D structures, has recently witnessed an accumulation of experimentally determined protein 3D structures whose functions are unknown (Berman, H.M., & Westbrook, J.D. 2004). This, in turn, has created the need for computational methods of structure-based protein function prediction, especially those which can be implemented automatically in high-throughput fashion (Jung, J.W. & Lee, W. 2004; Yakunin, A.F., et al., 2004). One of the main roles of bioinformatics (or computational biology) in this post-genomic era of biology is to reduce the workload of the experimentalists by computationally “eliminating” candidates for experimentation, thereby allowing them to invest their time and effort on the “good” ones that are more likely to yield useful results. This is one of the main objectives of the present work.

There are a number of established ways to predict (computationally) the function of a protein whose 3D structure (and amino acid sequence) is known. One way to do this is to predict the ligand(s) that the protein binds. To do this based on the 3D structure of the protein, one can proceed by detecting the ligand’s binding site - the ligand’s specific ‘signature’ on its receptor protein - in the receptor protein’s 3D structure. Ligands usually dock on the surface of a protein, and a ligand’s binding site (BS, LBS) is “buried” within the receptor protein’s interior to varying degrees.

The work we describe here involves the prediction of the binding sites of six biologically important ligands, namely: GTP, ATP, sialic acid, retinoic acid and nitric oxide in heme-bound and unbound forms. The biological roles of GTP and ATP are widely established (for example, see Mazzorana M, et al., 2008, and Stork PJ., 2003, respectively) Since both GTP- and ATP-binding proteins are highly heterogeneous, we focus here on the small Ras-type G-proteins (SRGP) and the ser/thr protein kinase (STPK) families, respectively. Sialic acid (SIA) is a C9 monosaccharide, and is the key component of mucus that allows the latter to prevent infections; more importantly, however, it has a significant role in the regulation of cellular communication (Lehmann et al., 2006; Miyagi et al., 2004). Retinoic acid (REA), on the other hand, has important roles in the transcriptional modulation of certain target genes by interacting with any one of its three known receptors: alpha, beta and gamma (Germain et al., 2006; Wolf, 2006). Finally, nitric oxide (NO) is an important signaling molecule in various cell types (Cary et al., 2006; Brunori et al., 2006; Perreti et al., 2006; Russwurm et al., 2004) which may either be in heme-bound (hNO) or unbound (fNO) forms.

The binding sites of these ligands were first characterized from ligand-containing experimentally solved structures from the PDB. These collection of structures from which the binding mode of the ligands are “learned” by an algorithm is called the ‘training set’ for the specific ligand in question. The salient features of the binding sites are then encoded in a tetrahedral tree data structure we designate as the ‘3D search motif’ (3D SM). Using a novel analytical screening algorithm we developed earlier (Reyes, V.M., & Sheth, V.N., 2011; Reyes, V.M., 2015a), a set of some 801 experimentally solved but
functionally unannotated protein structures from the PDB were screened for these 3D search motifs. Of the 801, we detected 8 putative SRGP GTP-binding proteins, 61 putative STPK ATP-binding proteins, 35 putative SIA-binding proteins, 132 putative REA-binding proteins, 33 putative hNO-binding proteins, and 10 putative fNO-binding proteins. These candidate proteins were then subjected to the “cutting plane” and “tangent sphere” methods (Reyes, V.M., 2015b) as a further validation step. This method is a way to assess the degree of burial of a local functional site such as a ligand-binding site in a protein. The validation depends on the putative structure having the same or similar depth of ligand binding site burial as those in the training structures. To our knowledge, this work is the first computational investigation that predicts the binding sites for GTP, ATP, SIA, REA, hNO and fNO from among functionally unannotated structures in the PDB, and further screens those proteins using information regarding the depth of burial of the bound ligand within its cognate receptor protein.

2 Datasets and Methods.

2.1 The Training Structures.

The screening method used here has been reported previously by us (Reyes, V.M., & Sheth, V.N., 2011; Reyes, V.M., 2015a). It requires the construction of a ‘3D search motif’ (3D SM) from a set of training structures, and is based on the geometry and architecture of the ligand binding site (LBS) in question. The 3D SM is essentially a ‘signature’ of the LBS in question and contains at least six quantitative and eight qualitative parameters which are all inputted into the algorithm to enable it to detect the said LBS. The training structures for ATP-binding STPK proteins and GTP-binding SRGP proteins have been described and discussed in detail previously (ibid.). The training structures used for the construction of the 3D SM for the SIA are 1JSN, 1JSO, 1W0O, 1W0P, 1MQN (chains A and D); the training structures used to construct the 3D SM for REA are 1FM9, 1K74, 1FBY (chains A and B), 1FM6 (chains A and U), 1XDK (chains A and E), 1XL5 (chains A, B, C and D), 2ACL (chains G, A, C and E); the training structures used for the construction of the 3D SM for hNO are 1OZW, 1XK3, 1ZOL (chains A and B); and finally, the training structures used to construct the 3D SM for fNO is 1ZGN, chains A and B. These training structures are all described in Table 1. The set of 801 protein structures in the PDB (all experimentally solved, mostly by x-ray crystallography) that lacked functional annotation at the time of this work are shown in Table 2. These proteins of unknown function come from many different species, but most are from E. coli, T. maritima, T. thermophilus, B. subtilis, P. aeruginosa, H. influenzae and A. fulgidus; only 18 (2.25%) come from H. sapiens. We used this set as the ‘application set’ – the set of 3D structures that we screened for the LBS’s in question for the purpose of assigning function to. In addition to determining the 3D SM from the above training structures, we also determined the depth of ligand burial in each, since this information is required in the next stages of our overall screening protocol.

2.2 Methods

2.2.1 Determination of the 3D SM’s.

The overall methodology followed in this work has been described in detail (Reyes, V.M., & Sheth, V.N., 2011; Reyes, V.M., 2015a). Briefly, the set of all hydrogen bonding (H-bonding) and van der Waals (VDW) interactions between ligand and protein in the training structures are sequestered (Engh, R.A. & Huber, R., 1991; Bondi, A., 1964). Then the most dominant and/or recurrent interactions among the training structures are determined, and designated the ‘3D binding consensus motif’. From such a consensus interaction mode between ligand and protein, the corresponding 3D SM is constructed. The 3D SM is a tetrahedral collection of four points in space representing the protein residues most commonly in association with the ligand (in the training structures). In the 3D SM, the
protein is in a reduced representation which we call the “double centroid reduced representation” (DCRR), where each amino acid is represented by two points, namely: the centroid of its backbone atoms (N, CA, C’, O), and that of its side chain atoms (CB, CG, etc.). The application set is then screened for the tetrahedral 3D SM using a screening algorithm we developed earlier (ibid.). The tetrahedral 3D SM’s for the six ligands in this study are shown in Figure 1, Panels A-D. The tetrahedral 3D SM is not just a collection of four points in space; it is a data structure that embodies a relatively large amount information about the binding site of the ligand in question. Specifically, it contains at least eight qualitative parameters, namely: the identities of the four amino acids in the tetrahedron (may be more if similar amino acids can interact with any of the ligand atoms in other receptor proteins) and their mode of association with the ligand (whether with backbone or side chain; hence, 4 x 2 = 8) and exactly six quantitative parameters (the lengths of the six sides of the tetrahedron) about the ligand binding site in question. Hence the 3D SM contains a total of 8 + 6 = 14 combined qualitative and quantitative parameters. This property makes the algorithm optimally specific for the ligand in question (ibid.).

2.2.2 Determining the Degree of Burial of the Ligand Binding Sites.

In our screening protocol, there are two further steps after the detection of the LBS’s using the 3D SM method (although this step is the most crucial). These two last steps depend on the “cutting plane” and “tangent sphere” methods (CPM and TSM, respectively) of ligand burial depth quantitative determination methods we reported previously (Reyes, V.M., 2015b; see also Figure 2). These two methods are complementary and produce numerical measures which we term the “CP index” (CPi) and “TS index” (TSi), respectively, and which are essentially quantitative measures of the degrees of burial of a given ligand or LBS. These two additional steps are meant to narrow down the set of structures testing positive for the presence of a particular LBS and thus serve to further validate the prediction results. Specifically, those which have LBS burial depths resembling those in the training structures are deemed more likely to be true positives than those whose degrees of LBS burial are quite different.

3 Results

The determination of the 3D SM and the validation stage (testing positive and negative control structures) for the GTP-binding site in SRGPs and the ATP-binding site in STPKs have both been presented and discussed in detail in our previous work (Reyes, V.M., 2015a), so we shall not touch upon them here and just limit our discussion in the following sections to SIA, REA, hNO and fNO binding sites.

3.1 Determination of the 3D SM for Sialic Acid (SIA) Binding Sites.

The H-bonds between SIA and its receptor protein in the training structures are dominated by interactions between atom N5 of SIA and the backbone O of a Gly or a Val residue in the BS; atom O1A of SIA and either an NH1 atom of an Arg or an OE1 or NE2 atom of a Gln residue in the BS; and atom O8 of SIA and the hydroxyl O of a Ser or a Tyr residue in the BS. The VDW interactions, on the other hand, are mainly between atom C7 of SIA and either the CH2 side chain atom of a Trp or the CE side chain atom of a Met residue in the BS. Careful consideration of these interactions enabled us to build the 3D search motif for SIA shown in Figure 1A.

3.2 Determination of the 3D SM for Retinoic Acid (REA) Binding Sites.

The H-bonds between REA and its receptor protein in the training structures are dominated by interactions between atom O1 of REA and the terminal side chain amino group of an Arg residue, and atom O2 of REA and the backbone N of an Ala residue in the BS. The VDW interactions, on the other hand, are mainly between atom C3 of REA and one of the side chain carbon atoms of an Ile or a Val
residue in the BS; atom C17 of REA and either atom CB of a His or a Cys residue, or the backbone O of a Cys residue in the BS; and finally atom C20 of REA and the CD2 atom of a Phe or a Leu residue in the BS. Careful consideration of these interactions led us to construct the 3D SM for REA shown in Figure 1B.

3.3 Determination of the 3D SM for Heme-Bound Nitric Oxide (hNO) Binding Sites.

The H-bonds between hNO and its receptor protein in the training structures are dominated by interactions between the heme iron and the side chain amino group of a His residue in the BS; the O2D atom of heme and a side chain amino group of an Arg residue in the BS; and atom O of NO and a Gly residue atom or a side chain C atom of a Leu residue in the BS. The VDW interactions, on the other hand, are mainly between atom CHD of heme and a side chain C atom of a Phe or a Gly residue in the BS; and between atom O2D of heme and a side chain atom of a His or a Tyr residue in the BS. Careful consideration of the above interactions allowed us to build the 3D SM for hNO shown in Figure 1C.

3.4 Determination of the 3D SM for Free/Unbound Nitric Oxide (fNO) Binding Sites.

The H-bonds between fNO and its receptor protein in the single training structure (with two protein chains) involve N atom of fNO and the backbone N of an Arg residue or the side chain OH group of a Tyr residue in the BS. The VDW interactions, on the other hand, are mainly between the N atom of fNO and an atom of a Gly or a Val residue in the BS, or between atom O of fNO and a side chain C atom of an Ile or Phe residue in the BS. Careful consideration of these interactions allowed us to construct the 3D SM for unbound fNO shown in Figure 1D.

3.5 Validation Step: Positive and Negative Controls

3.5.1. Negative Control Structures.

Thirty negative control structures were used for the validation of the BS’s for all six protein families studied here; they are, namely: 135L, 1A1M, 1A6T, 1BHC, 1PSN, 1BRF, 1EWK, 1CBN, 1MV5, 1JFF, 104M, 1ASH, 1B3B, 1BRF, 1CKO, 1CRP, 1EWK, 1F3O, 1FW5, 1HWY, 1JB, 1MJ, 1MV5, 1NQT, 1OGU, 1P6E, 1RDQ, 1SVS, 1TWY and 1Z3C. The above structures are all described in Table 3. Our results show that in all cases, the algorithm found no 3D SM in any of the negative control structures as expected. These results imply that the algorithm is highly specific for their respective ligands.

3.5.2. Positive Control Structures.

As for positive control structures, we note that there are no other appropriate positive structures in the PDB for the four above ligands as all of them have been used as training structures. Positive control structures to be used for validation must be yet “unseen” by the algorithm. We thus constructed artificial positive control structures from the negative control structures by replacing four appropriate amino acid residues in the latter to make a legitimate 3D SM for the particular ligand. These artificially mutated structures were then screened for the appropriate 3D SM using our algorithm. In all cases, the algorithm detected the artificially embedded 3D SM for the particular ligand (data not shown). These results imply that the screening algorithm has high sensitivity for the 3D SM corresponding to the particular ligand.

3.6 Screening Results

The screening process is illustrated in Figure 2. There are three stages in our screening process, the first stage and the most important being the LBS determination. The next stages involve the determination
of the LBS burial in the putative structures from the preceding stage. This is done by determining their CPi and TSi, respectively. The computed values are compared against the CPi and TSi of the respective training structures, and those putative structures having CPi and TSi closest to any of those of the training structures are deemed “double positives”; and are thus are considered best ligand-binding candidates in their respective protein families (see below). The application set, the 801 functionally unannotated structures in the PDB that served as application structures for this study, is shown in Table 2. These proteins come from a diverse distribution of species (see Table 4). The ‘Cutting Plane’ and ‘Tangent Sphere’ methods, on the other hand, are illustrated in Figure 1, Panels A and B, of our previous paper (Reyes, V.M., 2015b), which schematically illustrate the two methods and how they complement each other.

Overall results are as follows: of the 801 application structures, we detected 61 putative ATP-binding STPK proteins (7.6%), eight GTP-binding SRGP proteins (1.0%), 35 putative SIA-binding proteins (4.4%), 132 putative REA-binding proteins (16.5%), 33 putative hNO binding proteins (4.1%), and 10 fNO binding proteins (1.2%). We now show the details of these screening results in the following sections. Note that a protein that tested positive for a particular LBS may have more than one chain, and one or more LBSs may have been detected in each chain.

In the first 6 subtables of Table 5, the blue entries on top are the training structures for the particular 3D SM. Meanwhile, the black entries below are the structures that tested positive for the ligand in question. The headings “CPM” and “TSM” stand for “cutting plane” and “tangent sphere” methods, respectively. The red arrows point out those positive structures whose CPM and TSM indices are either within an arbitrarily set difference, e.g., within 8-10%, from any one of those of the training structures, respectively, of the closest one in the set. In each case, integration of these ligand burial depth results with those of the LBS screening results further trim down the positive set, at the same time further validating the LBS existence prediction. The information contained in the different parts of the tables are illustrated and explained diagrammatically in part 7 (of 7) of Table 5. Note that due to the large number of structures testing positive for the LBS (first stage of screening) in question in the two cases of ATP-binding STPK and REA-binding protein families (Table 5, part 2 of 7 and part 4 of 7, respectively) this diagram is not strictly adhered to. Instead, only structures with CPi and TSi values within 10.0 Å of those of a training structure are shown.

3.6.1 Screening Results for GTP-Binding Sites in Small Ras-type G-Proteins. Eight structures (1.0% of the original 801) tested positive in the initial screening step, the detection of the 3D SM for GTP (Table 5, part 1 of 7). This set then got reduced to seven (0.9%) after matching their CPi or TSi (i.e., at least one of them) values to those of the training structures. From these seven structures, two (0.2%) stand out, namely, 1XT1 and 1RU8, because both of their CPi and TSi values are close to those of one of the structures in the training set for this ligand (see Table 6).

3.6.2 Screening Results for ATP-Binding Sites in ser/thr Protein Kinases. The number of structures that tested positive for the ATP BS for this family is 61 (7.6%; see Table 5, part 2 of 7). By incorporating the ligand burial depth data from the CP and TS methods, 24 of the 61 structures testing positive for the ATP-binding site have been eliminated, leaving 37 structures (4.6%). Out of these 37, the following 11 to 13 structures (ca. 1.6%) are strong candidates because their CPi’s and TSi’s resemble both those of a training structure: 1WM6, 2CV1, 1RKQ, 1NF2, 1TQ6, 1MW0, 1TT7, 1T57, 1F19, 1RKi, 1Y9E (and possibly 1VPH and 1YYV as well; see Schwarzenbacher R. et al., 2004; Teplyakov A. et al., 2002; Beeby M. et al., 2005; Kunishima N. et al., 2005; see also Table 6).

3.6.3 Screening Results for Sialic Acid Binding Sites. For this ligand, 35 (4.4%) structures tested positive for the SIA binding site (Table 5, part 3 of 7). Of these, only 20 (2.5%) possess either a CPi or TSi close to that of a training structure. Of these 20, two structures (0.2%) namely, 1VKA and 1IU0, stand out as both of their CPi and TSi values resemble both the CPi and TSi values of one of the structures in the training set for this ligand (see Table 6).
3.6.4 Screening Results for Retinoic Acid Binding Sites. Of the 801 application structures, 132 (16.5%) tested positive for the REA binding site. Incorporating the ligand burial depth data from the CPM and TSM methods, almost 60% of the above 132 structures have been eliminated, leaving 53 candidate structures (6.6%; Table 5, part 4 of 7). The following 13 or 14 structures (ca. 1.7%) are strong candidates because both their CPI’s and TSi’s resemble both those of a training structure: 1YEY, 1NX4, 1TU1, 1Y8T, 2EVR, 1WU8, 1R1H, 1U61, 1T6S, 1NX8, 1NJH, 1Z6M, and 1VIM (and possibly 1ZE0 as well; see Clifton, I.J. et al., 2003; Asch WS, Schechter N., 2000; see also Table 6).

3.6.5 Screening for Heme-Bound NO Binding Sites. For this ligand, 33 structures (4.1%) tested positive for the hNO binding site. They have been further trimmed down to 12 (1.5%) upon incorporation of the ligand burial data using the CPM and TSM (Table 5, part 5 of 7). Of these 12, four structures (0.5%), namely 1ZSW, 1VKH, 1UAN and 2B4W stand out as their CPI and TSi values resemble both those from a training structure for this ligand (see Arndt, J.W. et al. 2005; Zhou C.Z. et al., 2005; see also Table 6).

3.6.6 Screening for Free/Unbound NO Binding Sites. In this set, the 10 structures (1.2%) tested positive for the fNO binding site. These have been narrowed down to six (0.7%) upon including the results from the CPM and TSM ligand burial data (Table 5, part 6 of 7). Of these six, a single structure (0.1%), namely, 1UC2, stands out as its CPI and TSi values both resemble those by the lone training structure, 1ZGN., for this ligand (see Table 6).

4 Discussion.

Using a novel analytical screening algorithm we developed earlier (Reyes, V.M., & Sheth, V.N., 2011; Reyes, V.M., 2015a), we have screened some 801 functionally unannotated x-ray diffraction structures deposited in the PDB for the binding sites of GTP, ATP, sialic acid, retinoic acid, and heme-bound and unbound nitric oxide. We detected eight SRGP GTP-binding sites, 61 STPK ATP-binding sites, 35 SIA-binding sites, 132 REA-binding sites, 33 hNO-binding sites and 10 fNO binding sites, with some structures containing more than one binding site for the ligand in question. The detection of the LBS for a particular ligand was accomplished by detecting the 3D SM for that ligand in the protein structures. This idea depends on the assumption that the 3D SM (and hence the binding site characteristics) for a given ligand is conserved within a protein family.

Using another novel analytical method we developed earlier (Reyes, V.M., 2015b) called the “cutting plane” and “tangent sphere” methods, the degrees of burial of these ligand binding sites were also determined and used as a further validation step for the ligand binding prediction. Thus the positive structures above were further culled by comparing their CPI or TSi to those of the training structures for the protein family and those which had similar values were retained, the rationale being those which have depths of LBS burial resembling those in the training structures are deemed more likely to be true positives than those who do not. This criterion depends on the reasonable premise that ligand burial depth is characteristic of a particular ligand-binding protein family.

Our LBS detection method depends on the availability of protein complex 3D structures with the bound ligand under study and as such relies heavily on the contents of the PDB. Although experimental structures for GTP- and ATP-binding proteins abound in the PDB, structures of proteins bound with other ligands are underrepresented. For example, the scarcity of structures containing SIA, REA, hNO and fNO in the PDB is a limitation in terms of having an ample number of both training and control (validation) sets for our screening algorithm. However, since our screening algorithm is largely analytical, the need for exhaustive positive and negative control structures is not that critical compared to statistical algorithms such as those based on SVM and neural networks. This is one advantage of an analytical algorithm over a stochastic one.

The fuzzy factor or margin, ε, we incorporate into the branches and node-edges in the 3D SM are usually in the order of 1.0 - 1.5 Å (Reyes, V.M. & Sheth, V.N., 2011; Reyes, V.M., 2015a). Thus in
cases where the protein assumes drastic conformational changes upon ligand binding and
displacements of amino acid residues at the binding site are much greater than 1.5 Å, our method will
perhaps likely fail. We believe it is reasonable to assume that the deeper within the protein interior the
LBS lies, the more drastic the conformational changes the protein undergoes upon binding the ligand
(i.e., in transitioning from the ‘apo’ to the complexed form). But whether or not the predictive power
of our algorithm decreases as the LBS lies deeper within the protein remains to be investigated.

In the determination of H-bonds between protein and ligand to build the 3D SM, we did not ascertain
the linearity of the bonds of the interacting atoms between ligand and protein (amino acids in the BS);
we merely measured non-hydrogen interatomic distances and we sequester only those with perfect or
near-perfect H-bond distances (2.7Å-2.9 Å). Thus this issue is unlikely to have a significant adverse
effect on our results, as instances in which the H-bonding atoms have perfect or near-prefix H-
bonding distances and at the same time non-linear, are quite rare.

5 Summary and Conclusions.

By determining the most prevalent and/or dominant H-bonding and VDW interactions between ligand
atoms and amino acid residue atoms in the BS of its receptor protein, we have constructed a ‘signature’
of the binding sites of six biologically important ligands – GTP, ATP, SIA, REA, hNO and fNO. We
designate this ‘signature’ as the 3D BS consensus motif for the particular ligand. We have then
encoded these binding site signatures in a tetrahedral tree data structure we call the 3D search motif or
“3D SM” for the ligand in question. Then, using a novel analytical search algorithm we developed
earlier (Reyes, V.M., & Sheth, V.N., 2011; Reyes, V.M., 2015a) experimentally determined protein
structures in the PDB that lacked functional annotation were screened for the above five ligands. We
detected eight structures with the GTP-binding site of the SRGP family, 61 structures with the ATP-
binding site of the STPK family, 35 structures with the SIA binding site signature, 132 with the REA’s,
33 with the heme-bound NO’s, and 10 with the free NO’s. The positive proteins above were further
subjected to validation by determining the depth of burial of their LBS’s using their CPi and TSi values
and comparing them to those of their training structures. Respectively seven, 37, 20, 12, and six of
the GTP-, ATP-, SIA-, REA-, hNO- and fNO-binding proteins had either their CPi or TSi close to
those of a retaining structure for the protein family. Of these, respectively two, 28 (of which 13 stand
out from the rest), two, 30 (of which 14 stand out from the rest), four and one of the GTP-, ATP-, SIA-
-, REA-, hNO- and fNO-binding proteins had both of their CPi and TSi close to those of a retaining
structure for the protein family. Thus by incorporating information about the depth of LBS burial in
the positive proteins from the 3D SM screening, they can be further narrowed down significantly for
increased confidence in the LBS prediction. At this point in the protein function prediction process, the
job of the bioinformaticist is usually done and the experimentalists take over. Thus we are currently
awaiting experimental verification of the results we report here. Our final results are shown in Table 6.

Acknowledgments. This work was supported by an Institutional Research and Academic Career
Development Award to the author, NIGMS/NIH grant number GM 68524. The author also wishes to
acknowledge the San Diego Supercomputer Center, the UCSD Academic Computing Services, and the
UCSD Biomedical Library, for the help and support of their staff and personnel. He also acknowledges
the Division of Research Computing at RIT, and computing resources from the Dept. of Biological
Sciences, College of Science, at RIT.

References

Arndt JW, Schwarzenbacher R, Page R, Abdubek P, Ambing E, Biorac T, Canaves JM, Chiu HJ, Dai
X, Deacon AM, DiDonato M, Elsliger MA, Godzik A, Grizzini C, Grzechnik SK, Hale J, Hampton E,
Han GW, Haugen J, Hornsby M, Klock HE, Koesema E, Kreusch A, Kuhn P, Jaroszewski L, Lesley SA, Levin I, McMullan D, McPhillips TM, Miller MD, Morse A, Moy K, Nigoghossian E, Ouyang J, Peti WS, Quijano K, Reyes R, Sims E, Spraggon G, Stevens RC, van den Bedem H, Velasquez J, Vincent J, von Delft F, Wang X, West B, White A, Wolf G, Xu Q, Zagnitko O, Hodgson KO, Wooley J, Wilson IA. "Crystal structure of an alpha/beta serine hydrolase (YDR428C) from Saccharomyces cerevisiae at 1.85 A resolution." Proteins. 2005 Feb 15;58(3):755-8.

Asch WS, Schechter N. "Plasticin, a type III neuronal intermediate filament protein, assembles as an obligate heteropolymer: implications for axonal flexibility." J Neurochem. 2000 Oct;75(4):1475-86.

Baxevanis, A.D. 2003. Using genomic databases for sequence-based biological discovery. Mol Med. 9:185-92.

Beeby M, O'Connor BD, Ryttersgaard C, Boutz DR, Perry LJ, Yeates TO. "The genomics of disulfide bonding and protein stabilization in thermophiles." PLoS Biol. 2005 Sep;3(9):e309. Epub 2005 Aug 23.

Bentley, S.D., and Parkhill, J. 2004. Comparative genomic structure of prokaryotes. Annu Rev Genet. 38:771-92.

Berman, H.M., and Westbrook, J.D. 2004. The impact of structural genomics on the protein data bank. Am J Pharmacogenomics. 4:247-52.

Bondi, A, 1964, “Van der Waals Volumes and Radii”, J. Phys. Chem., 68(3):441-451

Brunori M, Forte E, Arese M, Mastronicola D, Giuffrè A, Sarti P. "Nitric oxide and the respiratory enzyme." Biochim Biophys Acta. 2006 Sep-Oct;1757(9-10):1144-54. Epub 2006 May 13.

Burley, S.K. 2000. An overview of structural genomics. Nat Struct Biol. 7 Suppl:932-4.

Cary SP, Winger JA, Derbyshire ER, Marletta MA. "Nitric oxide signaling: no longer simply on or off." Trends Biochem Sci. 2006 Apr;31(4):231-9. Epub 2006 Mar 10.

Clifton IJ, Doan LX, Sleeman MC, Topf M, Suzuki H, Wilmouth RC, Schofield CJ. "Crystal structure of carbapenem synthase (CarC)." J Biol Chem. 2003 Jun 6;278(23):20843-50. Epub 2003 Feb 28.

Engh, R.A. & Huber, R. (1991) Accurate bond and angle parameters for X-ray protein structure refinement. Acta Cryst. (1991). A47, 392-400

Germain P, Chambon P, Eichelle G, Evans RM, Lazar MA, Leid M, De Lera AR, Lotan R, Mangelsdorf DJ, Gronemeyer H. "International Union of Pharmacology. LX. Retinoic acid receptors." Pharmacol Rev. 2006 Dec;58(4):712-25.

Heinemann, U. 2000. Structural genomics in Europe: slow start, strong finish? Nat Struct Biol. 7 Suppl:940-2.

Jung, J.W., and Lee, W. 2004. Structure-based functional discovery of proteins: structural proteomics. J Biochem Mol Biol. 37:28-34.

Kunishima N, Asada Y, Sugahara M, Ishijima J, Nodake Y, Sugahara M, Miyano M, Kuramitsu S, Yokoyama S, Sugahara M. "A novel induced-fit reaction mechanism of asymmetric hot dog thioesterase PAAI." J Mol Biol. 2005 Sep 9;352(1):212-28.

Lehmann F, Tiralongo E, Tiralongo J. "Sialic acid-specific lectins: occurrence, specificity and function." Cell Mol Life Sci. 2006 Jun;63(12):1331-54.

Mazzorana M, Pinna LA, Battistutta R. "A structural insight into CK2 inhibition." Mol Cell Biochem. 2008 Sep;316(1-2):57-62.
Miller, W., Makova, K.D., Nekrutenko, A., and Hardison, R.C.  2004. Comparative genomics.  *Annu Rev Genomics Hum Genet.*  5:15-56.

Miyagi T, Wada T, Yamaguchi K, Hata K.  "Sialidase and malignancy: a minireview."  *Glycoconj J.*  2004;20(3):189-98.

Murphy, W.J., Pevzner, P.A., and O'Brien, S.J.  2004. Mammalian phylogenomics comes of age.  *Trends Genet.*  20:631-9.

Norvell, J.C., and Machalek, A.Z.  2000. Structural genomics programs at the US National Institute of General Medical Sciences.  *Nat Struct Biol.*  7 Suppl:931.

Perretti M, D'Acquisto F. "Novel aspects of annexin 1 and glucocorticoid biology: intersection with nitric oxide and the lipoxin receptor."  *Inflamm Allergy Drug Targets.*  2006 Apr;5(2):107-14.

Reyes, V.M. & Sheth, V.N., "Visualization of Protein 3D Structures in 'Double-Centroid' Reduced Representation: Application to Ligand Binding Site Modeling and Screening", 2011, Handbook of Research in Computational and Systems Biology: Interdisciplinary Approaches, IGI-Global/Springer, pp. 583-598.

Reyes, V.M. (2015a) "An Automatable Analytical Algorithm for Structure-Based Protein Functional Annotation via Detection of Specific Ligand 3D Binding Sites: Application to ATP Binding Sites in ser/th Protein Kinases and GTP Binding Sites in Small Ras-type G-Proteins" (forthcoming; submitted to www.arXiv.org on Jan. 17, 2015); Abstract: Reyes, V.M., "Pharmacophore Modeling Using a Reduced Protein Representation as a Tool for Structure-Based Protein Function Prediction", J. Biomol. Struct. & Dyn., Book of Abstracts, Albany 2009: The 16th Conversation, June 16-20, 2009, Vol. 26 (6) June 2009, p. 873

Reyes, V.M. (2015b) "Two Complementary Methods for Relative Quantification of Ligand Binding Site Burial Depth in Proteins: The ‘Cutting Plane’ and ‘Tangent Sphere’ Methods" (forthcoming; submitted to www.arXiv.org on Feb. 5, 2015); Abstract: Cheguri, S. and Reyes, V.M., “A Database/Webserver for Size-Independent Quantification of Ligand Binding Site Burial Depth in Receptor Proteins: Implications on Protein Dynamics”, J. Biomol. Struct. & Dyn., Book of Abstracts, Albany 2011: The 17th Conversation, June 14-18, 2011, Vol. 28 (6) June 2011, p. 1013

Russwurm M, Koesling D. "NO activation of guanylyl cyclase."  *EMBO J.*  2004 Nov 10; 23(22):4443-50. Epub 2004 Oct 28.

Schwarzenbacher R, Deacon AM, Jaroszewski L, Brinen LS, Canaves JM, Dai X, Elsliger MA, Floyd R, Godzik A, Gritti C, Grzechnik SK, Klock HE, Koesema E, Kovarik JS, Kreusch A, Kuhn P, Lesley SA, McMullan D, McPhillips TM, Miller MD, Morse A, Moy K, Nelson MS, Ouyang J, Page R, Robb A, Quijano K, Spraggon G, Stevens RC, van den Bedem H, Velasquez J, Vincent J, von Delft F, Wang X, West B, "Crystal structure of a putative glutamine amido transferase (TM1158) from Thermotoga maritima at 1.7 A resolution."  *Proteins.*  2004 Mar 1;54(4):801-5.

Stork PJ. "Does Rap1 deserve a bad Rap?"  *Trends Biochem Sci.*  2003 May;28(5):267-75.

Teplyakov A, Obmolova G, Tordova M, Thanki N, Bonander N, Eisenstein E, Howard AJ, Gilliland GL.  "Crystal structure of the YjeE protein from Haemophilus influenzae: a putative ATPase involved in cell wall synthesis."  *Proteins.*  2002 Aug 1;48(2):220-6.

Terwilliger, T.C.  2000. Structural genomics in North America.  *Nat Struct Biol.*  7 Suppl:935-9.

Wolf G "Is 9-cis-retinoic acid the endogenous ligand for the retinoic acid-X receptor?"  *Nutr Rev.*  2006 Dec;64(12):532-8.

Yakunin, A.F., Yee A.A., Savchenko, A., Edwards, A.M., and Arrowsmith, C.H.  2004. Structural proteomics: a tool for genome annotation.  *Curr Opin Chem Biol.*  8:42-8.
Zhou CZ, Meyer P, Quevillon-Cheruel S, De La Sierra-Gallay IL, Collinet B, Graille M, Blondeau K, François JM, Leulliot N, Sorel I, Poupon A, Janin J, Van Tilbeurgh H. "Crystal structure of the YML079w protein from Saccharomyces cerevisiae reveals a new sequence family of the jelly-roll fold." Protein Sci. 2005 Jan;14(1):209-15.

FIGURE LEGENDS:

Figure 1. Panels A-D: The 3D Search Motifs. The 3D search motifs for the four ligands under study are shown: Panels A-D: Sialic acid, retinoic acid, heme-bound nitric oxide, and unbound nitric oxide search motifs, respectively. The lengths of the six sides of the tetrahedral motif (in Å) are also shown in an accompanying side table; the numbers inside parentheses are the corresponding standard deviations from the training structures. The ligand in each case is shown with its component atom names. The amino acids representing the tetrahedral vertices are indicated, with "(s)" indicating side chain interaction with ligand, and "(b)", backbone interaction. The root and three nodes are also indicated by the boxed letters.

Figure 2. The Elimination Process. Both local and global structure information are utilized in the process of elimination to search for candidate positive structures. Set A, the outermost red circle, represents the starting test/application set composed of 801 PDB structures without functional annotation. They are first screened for the particular 3D SM in question, and those that test positive, i.e., those that possess the 3D SM, form a subset of A; we call it set B (blue circle). Set B structures are then subjected to the “Cutting Plane” and “Tangent sphere” Methods (Reyes, V.M., 2015b). The CPM and TSM indices (CPMi and TSMi, respectively) of each structure are then compared respectively to those of the training structures used to create the 3D SM’s. Those whose CPMi or TSMi are within several units (typically 8-10) of those of the training structures, are considered to have similar indices, and form a subset of B; we call it set C (brown circle). Structures in set C are further analyzed to determine whether their indices are both respectively similar to those any one or more of the training structures. Those which satisfy this criterion form a subset of C, which we call set D (green circle). This main advantage of this elimination procedure is it can be automated and ran in batch or high-throughput mode, without the requirement for human intervention, a feature desired of analytical tools for large datasets.

TABLE LEGENDS:

Table 1. The Training Sets. The training structures for the determination of the 3D SMs for the binding sites of sialic acid, retinoic, and heme-bound and unbound nitric oxide, are shown. The PDB IDs of the structures are shown on column 1, the source organism on column 2, and a brief description of the structures is on column 3.

Table 2. The Control Structures. Negative control structures for SA, RA, hNO and fNO binding sites are shown. As for positive control structures for those ligand binding sites, please see text. Positive and negative control structures used for validating the 3D SM’s for GTP-binding SRGP and ATP-binding STPK protein families are taken up in detail in our previous work (Reyes, V.M., 2015a).

Table 3. The 801 Functionally Unannotated Proteins in the PDB Used As Application Structures. These 801 structures were obtained from the PDB in early 2006 by querying the PDB search site with the words “unknown function” or similar phrase. The absence of functional annotation in all 801 structures was further confirmed by examining the header information in each PDB file, which contained the phrase “function unknown” or a similar one in each case.
Table 4. Species Distribution of Application Set. The species distribution of the 801 application structures is shown in this table. The 801 structures come from 104 known species (that include bacteria, archaea, protozoans, and some higher organisms including humans), an uncultured bacterium (unknown species), and one is a synthetic protein. The 5 most represented species are *E. coli* (11.0%), *T. maritima* (7.9%), *T. thermophilus* (7.0%), *B. subtilis* (6.0%) and *P. aeruginosa* (4.7%).

Table 5, Parts 1-7. Cutting Plane and Tangent Sphere Indices Used to Assess LBS Burial in Screening Results: GTP in Small Ras-type GP (part 1 of 7); ATP in ser/thr PK; (part 2 of 7); Sialic Acid (part 3 of 7); Retinoic Acid (part 4 of 7); Heme-Bound NO (part 5 of 7); Unbound NO (part 6 of 7). The information in the above six tables is illustrated and identified schematically in part 7 of 7 of the table. Results of the determination of the particular binding site burial using the “cutting plane” and tangent sphere” methods are shown (headings “CPM” and “TSM”, respectively). The degree of burial is expressed as % of protein atoms on the exterior side of the cutting plane and inside the tangent sphere, respectively. Part 7 of 7 diagrammatically identifies what information are contained in the tables above based on their location in the table.

Table 6. Application Structures that Tested Positive. The structures from the set of 801 functionally unannotated proteins (Table 3) in the PDB that tested positive of the 3D SMs of GTP (in small, Ras-type G-proteins, ATP (in ser/thr protein kinases), sialic acid, retinoic acid, and heme-bound and unbound nitric oxide are summarized in this table. Note that most of the structures are still functionally unannotated at the time of this writing, as shown by the scarcity of entries in the last column, which is the published reference papers for the particular structure (see also References section).
FIGURES:

Figure 1A.
Figure 1B.
3D Search Motif #1: Heme-bound Nitric Oxide:

![Diagram of heme-bound nitric oxide]

| side | length, Å   | uncertainty, Å |
|------|-------------|----------------|
| Rn1  | 8.180       | 0.098          |
| Rn2  | 9.075       | 0.086          |
| Rn3  | 7.911       | 0.093          |
| n1n2 | 5.724       | 0.084          |
| n1n3 | 9.915       | 0.142          |
| n2n3 | 7.988       | 0.202          |

Figure 1C.
3D Search Motif #2: Non-heme Nitric Oxide:

![Diagram of 3D structure with atoms labeled]

| side | length, Å   | std. dev. |
|------|-------------|-----------|
| Rn1  | 5.096 (0.099) |
| Rn2  | 10.706 (0.073) |
| Rn3  | 6.736 (0.131) |
| n1n2 | 8.592 (0.022) |
| n1n3 | 3.562 (0.003) |
| n2n3 | 7.996 (0.016) |

Figure 1D.
Figure 2.

Set D: those whose TSMi and CPMi are both similar to that of (a) specific training structure/s

Set C: those whose TSMi or CPMi is similar to that of one of the training structures

Set B: those where at least one 3D SM was detected

Set A: the 801 PDB structures with unknown function

The Elimination Process Employed
TABLES:

The Training Sets

| Sialic Acid Binding Site | Protein or Virus                     | Description                                                                 |
|-------------------------|--------------------------------------|-----------------------------------------------------------------------------|
| 1JSN:A                  | influenza A virus                    | Hemagglutinin HA1 chain (residues 1-325, chain A) and HA2 chain (residues 1-170, chain B) with bound N-acetyl-D-glucosamine (NAG), D-galactose (GAL) and 0-sialic acid (SIA) |
| 1JCO:A                  | influenza A virus                    | Hemagglutinin HA1 chain (residues 1-325, chain A) and HA2 chain (residues 1-170, chain B) with bound N-acetyl-D-glucosamine (NAG) and 0-sialic acid (SIA) |
| 1W00:A                  | Vibrio cholerae                      | Saltase (EC 3.2.1.16; syn.: neuraminidase, nanase) with bound calcium ion, 2-deoxy-2,3-didehydro-2-deoxy-p-glucosaminic acid (DBN) and 0-sialic acid (SIA) |
| 1W0P:A                  | Vibrio cholerae                      | Saltase (EC 3.2.1.16; syn.: neuraminidase, nanase) with bound calcium ion, glycerol (GL), 2-amino-2-hydroxyethoxypropane-1,3-diol (TRIS), and 0-sialic acid (SIA) |
| 1MQN:A,D                | influenza A virus                    | Hemagglutinin HA1 chain (chains A, D, G) and HA2 chain (chains B, E, H) with bound N-acetyl-D-glucosamine (NAG), alpha-D-mannose (MAN), D-galactose (GAL) and 0-sialic acid (SIA) molecules |

Retinoic Acid Binding Site:

| Protein               | Species          | Description                                                                 |
|-----------------------|------------------|-----------------------------------------------------------------------------|
| 1FM9:A                | Homo sapiens     | The heterodimer of the human RXR-alpha and PPAR-gamma ligand binding domains respectively bound with 9-cis retinoic acid and G262070 and co-activator peptides |
| 1K74:A                | Homo sapiens     | The heterodimer of the human PPAR-gamma and RXR-alpha ligand binding domains respectively bound with GW40644 and 9-cis retinoic acid and co-activator peptides |
| 1FBy:A,B              | Homo sapiens     | The human RXR-alpha ligand binding domain bound to 9-cis retinoic acid |
| 1FM6:A,U              | Homo sapiens     | The heterodimer of the human RXR-alpha and PPAR-gamma ligand binding domains respectively bound with 6-cis retinoic acid and resiglitazone and co-activator peptides |
| 1XDK:A,E              | Mus musculus     | The RAR-beta/RXR-alpha ligand binding domain heterodimer is complex with 9-cis retinoic acid and a fragment of the TRAP220 co-activator |
| 1XLS:A,B,C,D          | Homo sapiens /   | Heterodimer of the human RXR-alpha ligand binding domain and the mouse orphan nuclear receptor LRH (syn.: constitutive androstane receptor, CAR) bound to TCPBP3 and 9-cis retinoic acid and a TIP2 peptide containing the 3rd LXXL motif |
| 2ACL:A,C,E,G          | Homo sapiens     | Heterodimer of the human retinoic acid RAR-alpha and the mouse orphan receptor LXR-alpha (syn.: nuclear orphan receptor LXR-alpha) with bound 1-benzyl-3,4-methoxyphenylaminol4-phenylpiperole-7,8-dione (LBD) and retinoid acid |

Heme-Bound Nitric Oxide:

| Protein               | Species          | Description                                                                 |
|-----------------------|------------------|-----------------------------------------------------------------------------|
| 1OZW:A,B              | Homo sapiens     | The fumaric, ferric and ferrous-nitrogen oxide (heme-complexed) forms of the human heme oxygenase-1 (EC 1.14.99.3) |
| 1XK3:A,B              | Homo sapiens     | Heme oxygenase-1 (H.O.1) Arg163Glu mutant with bound heme-complexed nitrogen oxide (NO) |
| 1OZL:A,B              | Homo sapiens     | The fumaric, ferric and ferrous-nitrogen oxide (heme-complexed) forms of the Asp460Ala mutant of human heme oxygenase-1 (EC 1.14.99.3) |

Unbound Nitric Oxide:

| Protein               | Species          | Description                                                                 |
|-----------------------|------------------|-----------------------------------------------------------------------------|
| 1ZGN:A,B              | Homo sapiens     | Glutathione-s-transferase pi (syn.: GST Class P) with bound dinitrosyl-diglutathionyl iron complex |

Table 1.
Table 2 (part 1 of 2).

| 1DI6 | 1J3L | 1LJ7 | 1NNQ | 1O6A | 1QVV | 1RE2 | 1SYR | 1TU9 | 1V70 |
|------|------|------|------|------|------|------|------|------|------|
| 1DT7 | 1JAL | 1LJO | 1NNW | 106D | 1QYV | 1RE3 | 1T66 | 1TUA | 1V6D |
| 1DM5 | 1JH1 | 1LPL | 1NXI | 1089 | 1QZV | 1S12 | 1T07 | 1TUH | 1V9H |
| 1DW1 | 1J00 | 1LQ1 | 1NQ5 | 108C | 1QW2 | 1S2X | 1T0B | 1TVV | 1V80 |
| 1F99 | 1JQG | 1LXJ | 1NOG | 10NO | 1QY9 | 1S4C | 1T0T | 1TWW | 1VEP |
| 1F9L | 1J0P | 1LXN | 1NPD | 10Q1 | 1QA4 | 1S4K | 1T11 | 1WU | 1V96 |
| 1FX1 | 1J0V | 1M1S | 1NYY | 10QU | 1QY1 | 1S5A | 1T2B | 1WY | 1V99 |
| 1G2R | 1JKI | 1M33 | 1NQM | 10SC | 1Q24 | 1S7H | 1T3U | 1XJ | 1V9B |
| 1H2H | 1JRK | 1M3S | 1NQN | 10Y1 | 1Q28 | 1S7I | 1T57 | 1XXL | 1VAJ |
| 1HQQ | 1J5X | 1M65 | 1NR9 | 1OY2 | 1R0U | 1S7O | 1T5J | 1XZ | 1VEK |
| 1HRU | 1J5X | 1M63 | 1NRI | 1OZ9 | 1R3D | 1SNW | 1T5R | 1T70 | 1VEV |
| 1HTW | 1J3H | 1M34 | 1NRR | 1P1L | 1R47 | 1SN9 | 1T5Y | 1T70 | 1VC4 |
| 1HXL | 1J2T | 1MK4 | 1NS5 | 1P1M | 1R5X | 1SAW | 1T62 | 1T2A | 1V0H |
| 1HZX | 1J26 | 1ML8 | 1NU9 | 1P5P | 1R6Y | 1SBK | 1T6A | 1T2Z | 1V0W |
| 1HY2 | 1J2E | 1MOG | 1NX4 | 1P8C | 1R75 | 1SC0 | 1T6S | 1U05 | 1VE3 |
| 1IS6 | 1J3R | 1MW5 | 1NX8 | 1P99 | 1R7L | 1SD5 | 1T6T | 1U0K | 1VGG |
| 1IS0 | 1K4N | 1MW7 | 1NXH | 1P9I | 1RC6 | 1SDI | 1T8H | 1U5W | 1V0Y |
| 1IGN | 1K77 | 1NMQ | 1NXJ | 1P9Q | 1RCU | 1SDJ | 1T95 | 1U61 | 1VH0 |
| 1I5H | 1K7J | 1MW | 1NZ2 | 1PB0 | 1RFE | 1SE0 | 1T9F | 1U69 | 1VH5 |
| 1IHN | 1K7K | 1NZ3 | 1NY1 | 1PBJ | 1RFE | 1SEF | 1TC5 | 1U6L | 1VH6 |
| 1JJ6 | 1K9P | 1N1Q | 1NYE | 1P26 | 1R16 | 1SF9 | 1T6D | 1U73 | 1VH9 |
| 1IIV | 1KJN | 1N51 | 1MZA | 1PD3 | 1RKT | 1SPN | 1T6M | 1U7W | 1VHC |
| 1IN0 | 1KK9 | 1NC5 | 1NZJ | 1PF5 | 1RQ9 | 1SPS | 1T6L | 1U84 | 1VHE |
| 1IUJ | 1KON | 1NC7 | 1NZN | 1PG6 | 1RLH | 1SPX | 1T6L | 1U9C | 1VF0 |
| 1IUK | 1KQ5 | 1NE2 | 1Q01 | 1PM3 | 1RLJ | 1SG9 | 1T6Q | 1U9D | 1VEK |
| 1IUL | 1KQ4 | 1NE3 | 1Q13 | 1PQY | 1RLK | 1SH8 | 1TOO | 1U9P | 1VHN |
| 1IXL | 1KX4 | 1NF2 | 1Q51 | 1PT5 | 1RTT | 1SSH | 1TOO | 1UAN | 1VH0 |
| 1IXM | 1KXU | 1NG6 | 1Q22 | 1P27 | 1RTW | 1S05 | 1TOY | 1UC2 | 1VHE |
| 1IY7 | 1KXH | 1NJ9 | 1Q3U | 1P78 | 1R7Y | 1SMB | 1TP6 | 1UCR | 1VFQ |
| 1IJR | 1KYT | 1NI4 | 1Q4T | 1P06 | 1RU8 | 1SPV | 1TPX | 1UR8 | 1VHS |
| 1IJS | 1KBP | 1N1L | 1QAW | 1PV5 | 1RV9 | 1SQ4 | 1TQ5 | 1UFP | 1VHU |
| 1IK3 | 1LS | 1NHH | 1Q50 | 1PYM | 1RVK | 1SQE | 1TQ8 | 1UF9 | 1VHY |
| 1IKM | 1LSX | 1NKJ | 1Q51 | 1PW5 | 1RW0 | 1SQH | 1TQB | 1UPA | 1V11 |
| 1IKN | 1L6R | 1NJR | 1Q5J | 1Q2Y | 1RW1 | 1SQS | 1TCQ | 1UPB | 1V13 |
| 1IU5 | 1LGV | 1NKQ | 1Q5U | 1Q4R | 1R7W | 1SQV | 1TQX | 1UFO | 1V14 |
| 1I74 | 1LQW | 1NKV | 1Q61 | 1Q77 | 1RXD | 1SQW | 1TSJ | 1U98 | 1V17 |
| 1I7D | 1LQZ | 1NN | 1Q62 | 1Q7H | 1RXH | 1SR0 | 1T74 | 1UNJ | 1V18 |
| 1I9B | 1LOD | 1NMO | 1Q65 | 1Q8B | 1RXJ | 1SS4 | 1T77 | 1V30 | 1V1M |
| 1I9I | 1LQ | 1NMP | 1Q67 | 1Q8C | 1RXK | 1SU0 | 1T7Z | 1V6R | 1TV |
| 1J8K | 1LEL | 1NHH | 1Q69 | 1Q9U | 1RYL | 1SU1 | 1T0 | 1V6T | 1V12 |
Table 2  (part 2 of 2).

| 1VJ1 | 1VQW | 1VW8 | 1XKF | 1Y7I | 1YOA | 12C6 | 2A1V | 2ABZ | 2CWW |
| 1VJ2 | 1VQY | 1VW9 | 1XKL | 1Y7M | 1YOC | 12CE | 2A2L | 2ASF | 2CV9 |
| 1VJF | 1VQZ | 1VWI | 1XQK | 1Y7P | 1YOK | 12D0 | 2A2M | 2ATR | 2CVB |
| 1VJG | 1VR4 | 1WM5 | 1Y7R | 1YOM | 12D0 | 2A2O | 2ATZ | 2CVE |
| 1VKJ | 1VR9 | 1WNP | 1YM7 | 1Y00 | 1YOZ | 12DR | 2A33 | 2AU0 | 2CVL |
| 1VJM | 1W6I | 1Y6W | 1XM5 | 1Y82 | 1YQF | 12K2 | 2A3N | 2AVW | 2CW4 |
| 1VJO | 1WSA | 1X5I | 1XN4 | 1Y86 | 1YQH | 12RE | 2A3Q | 2AV4 | 2CWQ |
| 1VK0 | 1WD5 | 1X5J | 1XP7 | 1Y99 | 1YRE | 12KI | 2A5E | 2AVN | 2CYW |
| 1VK1 | 1WD6 | 1X72 | 1XQ4 | 1Y8A | 1YS9 | 12KP | 2A67 | 2AX3 | 2CX0 |
| 1VK5 | 1WDI | 1X77 | 1XQ6 | 1Y8T | 1YTL | 12L0 | 2A6B | 2AX0 | 2CX1 |
| 1VK8 | 1WDJ | 1X7F | 1XQ9 | 1Y9B | 1YTD | 12MB | 2A6C | 2AXP | 2D2Y |
| 1VK9 | 1WDT | 1X7V | 1XQA | 1Y9E | 1YVE | 12N6 | 2A8E | 2AZ4 | 2B4R |
| 1VKA | 1WDV | 1X9G | 1XQB | 1Y9I | 1YWL | 12N8 | 2A9F | 2AZP | 2E59 |
| 1VKB | 1WEL | 1XAO | 1XR6 | 1YAC | 1YWD | 12OX | 2A9S | 2BA0 | 2E6H |
| 1VCD | 1WEP | 1XAP | 1XRI | 1XAV | 1YWF | 12P6 | 2AAM | 2EB0 | 2E7N |
| 1VEH | 1X0 | 1X5J | 1X5V | 1YD2 | 1XK1 | 12FV | 2AB0 | 2BD0 | 2ETD |
| 1VEI | 1XJ0 | 1XBP | 1XTL | 1YB3 | 1YYV | 12PW | 2AB1 | 2BO0 | 2ETH |
| 1VMO | 1W2K | 1X5V | 1XTM | 1YRM | 1YZV | 12FY | 2ACA | 2BY1 | 2ETS |
| 1VNW | 1W4K | 1XW | 1XT0 | 1YRX | 1ZY | 12Q7 | 2BEG | 2BP0 | 2EUC |
| 1VUO | 1WCK | 1X8X | 1XUV | 1YCD | 1YZZ | 12S7 | 2ARU | 2BDZ | 2EUI |
| 1VL4 | 1WLU | 1X9Y | 1XV2 | 1YC | 1ZOF | 12SO | 2AVY | 2BE0 | 2EVE |
| 1VL5 | 1WLV | 1XBD | 1XS | 1YDF | 1ZIS | 12SW | 2AFC | 2BE3 | 2EVR |
| 1VL7 | 1WLE | 1XCC | 1XNO | 1YDH | 1Z40 | 12TC | 2AH5 | 2BEM | 2BV |
| 1VL8 | 1WME | 1XD1 | 1XNM | 1YDM | 1Z47 | 12TD | 2AH6 | 2BEN | 2BW0 |
| 1VM0 | 1WNN | 1XJ1 | 1X7 | 1YDW | 1ZGM | 12TP | 2AJ4 | 2B41 | 2ECW |
| 1VMF | 1WNS | 1X2 | 1XXL | 1YB5 | 1Z6N | 12TV | 2AJ2 | 2BA0 | 2EWR |
| 1VNH | 1WN9 | 1X58 | 1XY7 | 1YRM | 127A | 12UP | 2AJ6 | 2BW2 | 2F06 |
| 1VUJ | 1WNA | 1XI | 1XJF | 1Y0H | 1Y87 | 12VJ | 2AJ7 | 2B6C | 2F20 |
| 1VU2 | 1WOL | 1X5J | 1YUK | 1Y9 | 1Z84 | 12WJ | 2AL1 | 2B6E | 2F22 |
| 1VU4 | 1WQ | 1XFS | 1Y0N | 1YHF | 1Z55 | 12WY | 2AMH | 2BOM | 2F4L |
| 1VU6 | 1WED | 1XG7 | 1Y0Z | 1YKW | 1Z6H | 12X3 | 2AMU | 2EDE | 2F4N |
| 1VUB | 1WRE | 1XG8 | 1Y12 | 1YLK | 1Z50 | 12X5 | 2A99 | 2EDT | 2PF4 |
| 1VUD | 1WSC | 1XH1 | 1XYL | 1YLM | 1Z94 | 12X8 | 2AP3 | 2EDV | 2F9C |
| 1VUP | 1WTY | 1XH0 | 1Y2I | 1YLN | 1Z9T | 12XJ | 2AP6 | 2EB4 | 2FEL |
| 1VVP | 1WQ | 1XH6 | 1Y5 | 1YLN | 1ZBM | 12XO | 2AP7 | 2C6Q | 2FBM |
| 1VW | 1WD | 1XH8 | 1Y63 | 1YLO | 1ZB0 | 12XU | 2APL | 2COH | 2FDS |
| 1VP | 1WUF | 1XIL | 1Y6Z | 1YIX | 1ZB2 | 12ZN | 2AQW | 2CSL | 2FE1 |
| 1VQ | 1WS | 1XJC | 1Y71 | 1YN4 | 1ZBR | 2A13 | 2AR1 | 2C10 | 2FF9 |
| 1QS | 1WV3 | 1XK8 | 1Y7H | 1YN5 | 1ZBS | 2A15 | 2A85 | 2CO6 | 2FPI |

20
| PDB ID | DESCRIPTION |
|--------|-------------|
| 1AMC  | Sperm whale (Physeter catodon) skull SH3 domain (heme-iron site) bound with bound N-ethylmaleimide and sodium at pH 7.0 |
| 1ASH  | Structure of N-3-protoporphyrin (N3P) bound hemoglobin domain 1 from Acacia auriculiformis with bound dihydroxynaphthoquinone at 2.2 Å resolution |
| 1E3R  | Structure of glutamate dehydrogenase from Thermotoga maritima with mutations N272D and Q276K |
| 1BRF  | Structure of Puromycin with bound Fe(II) from Pyrococcus furiosus at 0.95 Å resolution |
| 1CEN  | Structure of the hydrophobic protein cocrystal from the seed of Camellia sinensis (Bourkajahim cabage) at 1.2 Å and at 0.83 Å resolution |
| 1CLO  | Structure of mRNA capping enzyme from Chloralovirus PBCV-1 in complex with the CAF analog OppOpG |
| 1CFP  | NMR structure (6K20) of human C-H-Ras p21 protein (catalytic domain, N1-100) complexed with GDP and MEK |
| 1CWW  | Structure of the metabotropic glutamate receptor subtype 1 from Rat functional expressed with glutamate |
| 1F2Q  | Structure of M3T95 ATP-binding cassette with bound Mg-ADP from Methanococcus jannaschii |
| 1FW5  | Solution structure of mammalian binding protein of Semliki forest virus mRNA capping enzyme NSP1 |
| 1HWW  | Glutamate dehydrogenase from Bos taurus complexed with NAD and Z-sulfooxinate |
| 1JF7  | Refined structure of bovine (Bos taurus) a-3 tubulin from zinc-induced sheets stabilized with tetracetic acid |
| 1KJ1  | Structure of the complex of the Ras bound to antitoxin antibody M58-12 and the transition-state analog, N3[1-14-carboxytylaminomethyl]-2-phenylethylamine-4-oxo]-3-acetyl-2-phenylethylamine |
| 1MV5  | Structure of the ATP-binding domain of the multidrug resistance ABC transporter and permarasa protein from Lactococcus lactis with bound ADP, ATP and Mg ion |
| 1NGT  | Structure of glutamate dehydrogenase from Bos taurus with bound ADP |
| 1PES  | Structure of papain (E.C.3.2.2.2) from the papaya fruit (Carica papaya) lipase complexed with E-64 (2S,2S)-3-[(3-methylbutyl)amino]-1-1H-1,3-carboxylate (oxo)]-2-carboxylate at 2.1 Å resolution |
| 1QG7  | Mycobacterium tuberculosis FT32 (ferritin temperature-sensitive mutant Z) in complex with ODP |
| 1SVS  | Structure of the K120F mutant of Glu subunit bound to GDPNH (phosphoanhydridic phosphoric acid guanylate ester) |
| 1TUB  | Electron diffraction structure of Stz acrole (p1) tubulin a-3 dimer with bound STP. ODP and taxolere |
| 1WYY  | Structure of a hypothetical ABC-type phosphatase transporter from Vibrio cholerae 01 Bovar strain |
| 1ZDC  | mRNA cap guanine-N7 methyltransferase from the mouse (Mus musculus) complexed with AcylAdoMet |

*These negative control structures were used for beta acids, catalase, and heme-bound and unbound native acids. Negative control structures used for ATP (serine protein kinases) and GTP (small, Ras-like Q-proteins) are enumerated and discussed in Reyes, V.M. 2008a.

Table 3.
### Table 4.

**Species Distribution of the 801 Application Structures**

| Source organism                        | No. structures | % of Total | Source organism                        | No. structures | % of Total |
|----------------------------------------|----------------|------------|----------------------------------------|----------------|------------|
| Escherichia coli                       | 88             | 10.50      | Methanobrevibacter bryantii            | 03             | 0.37       |
| Thermotoga maritima                   | 66             | 7.06       | Vibrio parahaemolyticus                | 02             | 0.25       |
| Thermus thermophiles                   | 50             | 6.05       | Sulfolobus tokodaii                    | 02             | 0.25       |
| Bacillus subtilis                      | 48             | 5.70       | Streptococcus mutans                   | 02             | 0.25       |
| Pseudomonas aeruginosa                 | 38             | 4.74       | Salmonella enterica                    | 02             | 0.25       |
| Haemophilus influenzae                 | 26             | 3.08       | Rattus norvegicus                      | 02             | 0.25       |
| Archaeoglobus fulgidus                | 23             | 2.77       | Pseudomonas putida                     | 02             | 0.25       |
| Pyroccocus furiosus                    | 20             | 2.36       | Plectomonas yeillii                    | 02             | 0.25       |
| Pyrococcus horikoshii                  | 16             | 1.95       | Neisseria meningitidis                 | 02             | 0.25       |
| Arachneoptera thailandica              | 10             | 1.20       | Methanomaculum nitriiphilum            | 02             | 0.25       |
| Staphylococcus aureus                  | 10             | 1.20       | Eubacterium rosea                      | 02             | 0.25       |
| Novus sapine                          | 10             | 1.20       | Desulfovoccus mediterraneus            | 02             | 0.25       |
| Bacillus alebei                        | 17             | 2.07       | Chromobacterium violaceum              | 02             | 0.25       |
| Mycobacterium tuberculosis            | 14             | 1.63       | Caldicella crecentis                   | 02             | 0.25       |
| Vibrio cholerae                       | 13             | 1.55       | Copia hircus                           | 02             | 0.25       |
| Enterococcus faecalis                  | 13             | 1.55       | Sorotella bronchiseptica               | 02             | 0.25       |
| Bacillus subtilis                      | 10             | 1.19       | Bacterium anthracis                    | 02             | 0.25       |
| Agrobacterium tumefaciens              | 10             | 1.19       | Unidentified bacteria                   | 01             | 0.12       |
| Streptomyces avidinii                  | 12             | 1.42       | Trypanosoma brucei                     | 01             | 0.12       |
| Thermoplasma acidiphilum               | 11             | 1.33       | Trypanosoma cruzi                      | 01             | 0.12       |
| Leishmania major                       | 10             | 1.19       | Toxoplasma gondii                      | 01             | 0.12       |
| Bacteroides thetaiotaomicron           | 10             | 1.19       | Synthesis protein                      | 01             | 0.12       |
| Streptomyces pyogenes                  | 08             | 0.98       | Synthesis glasii                       | 01             | 0.12       |
| Streptococcus pneumoniae               | 08             | 0.98       | Schizosaccharomyces pombe              | 01             | 0.12       |
| Salmonella typhimurium                 | 08             | 1.00       | Carys coronaviruses                    | 01             | 0.12       |
| Pyrococcus acolyphilum                 | 08             | 1.00       | Rhodopseudomonas palustris             | 01             | 0.12       |
| Plasmodium falgentz                  | 07             | 0.87       | Pseudomonas erythrae                   | 01             | 0.12       |
| Nitrospora europae                     | 07             | 0.87       | Flavobacterium berghei                | 01             | 0.12       |
| Methanobacterium thermautotrophicum    | 07             | 0.87       | Flavobacterium pivovar                 | 01             | 0.12       |
| Desulfovibrio desulfovestimentis       | 07             | 0.87       | Flavobacterium vulgari                 | 01             | 0.12       |
| Bacillus halodurans                    | 07             | 0.87       | Flavobacterium vulgari                 | 01             | 0.12       |
| Shewanella oneidensis                  | 06             | 0.73       | Mycoplasma genitalium                  | 01             | 0.12       |
| Methanococcus jannaschii               | 06             | 0.73       | Mycoplasma thermolestica               | 01             | 0.12       |
| Helicobacter pylori                    | 04             | 0.48       | Listeria monocytogenes                 | 01             | 0.12       |
| Aquifex aerolicola                     | 04             | 0.48       | Listeria innocua                       | 01             | 0.12       |
| Sulfolobus tokodaii                    | 03             | 0.37       | Leishmania donovani                    | 01             | 0.12       |
| Shigella flexneri                      | 03             | 0.37       | Listeria pneumos Tollera               | 01             | 0.12       |
| Vibrio arcticum                       | 03             | 0.37       | Listeria arctiva                      | 01             | 0.12       |
| Helicobacter pylori                    | 02             | 0.25       | Mycoplasma genitalium                  | 01             | 0.12       |
| Methanobacterium bryantii              | 02             | 0.25       | Mycoplasma thermautotrophicum          | 01             | 0.12       |
| Gallus gallus                          | 02             | 0.25       | Moraxella bovigenes                    | 01             | 0.12       |
| Capnocytophaga jejuni                  | 02             | 0.25       | Moraxella bovigenes                    | 01             | 0.12       |
| Porphyromonas gingivalis               | 02             | 0.25       | Moraxella bovigenes                    | 01             | 0.12       |
| Neisseria meningitidis                 | 02             | 0.25       | Moraxella bovigenes                    | 01             | 0.12       |
| Pyrococcus acolyphilum                 | 02             | 0.25       | Sorotella bronchiseptica               | 01             | 0.12       |
| Escherichia coli                       | 02             | 0.25       | Sorotella bronchiseptica               | 01             | 0.12       |
| Clostridium thermocellum               | 02             | 0.25       | Sorotella bronchiseptica               | 01             | 0.12       |
| Clostridium thermocellum               | 02             | 0.25       | Sorotella bronchiseptica               | 01             | 0.12       |
| Chlamydia sp.                          | 02             | 0.25       | Sorotella bronchiseptica               | 01             | 0.12       |
GTP-binding small Ras-type G-proteins

| CPM       | TSM       |
|-----------|-----------|
| 103y:A    | 8.0330    | 1m7b:A    | 16.6906    |
| 1nvu:C    | 10.2041   | 1e95:A    | 17.7361    |
| 1n61:A    | 10.3577   | 1loo:A    | 23.0242    |
| 2rap:     | 10.6095   | 1n61:A    | 27.6453    |
| 1m7b:A    | 16.4748   | 1nvu:Q    | 30.6122    |
| 1e96:A    | 16.5826   | 2rap:     | 31.3017    |
| 1loo:A    | 16.7015   | 1o3y:A    | 31.6817    |

|         |           |           |
| 1vhq:B  | 2.6710    | 2b30:C    | 2.4272     |
| * 1vhq:A| 3.0928    | 2b30:D    | 2.4713     |
| * 1oy1:D| 3.2967    | 2b30:B    | 2.5652     |
| * 1vim:C| 3.3058    | 2b30:A    | 2.6979     |
| * 1oy1:C| 3.4216    | * 1xtl:B  | 16.3979    |
| * 1oy1:E| 4.3478    | * 1ru8:A  | 19.4886    |
| * 1sg9:B| 5.6351    | * 1ru8:B  | 19.4886    |
| * 1oy1:A| 5.6747    | 1vim:C    | 42.4931    |
| * 1to0:B| 6.0124    | 1sg9:B    | 45.7737    |
| * 1to0:G| 6.3660    | 1oy1:A    | 51.0719    |
| * 1ru8:A| 7.5000    | 1oy1:B    | 53.7084    |
| * 1ru8:B| 7.5000    | 1oy1:D    | 56.2379    |
| * 1to0:D| 8.5963    | 1oy1:C    | 56.6172    |
| * 1xtl:B| 17.7817   | 1vhq:A    | 56.8299    |
| 2b30:A   | 25.6966   | 1vhq:B    | 58.3713    |
| 2b30:B   | 25.9620   | 1to0:D    | 66.5221    |
| 2b30:C   | 26.3019   | 1to0:G    | 75.5004    |
| 2b30:D   | 26.3019   | 1to0:B    | 76.7462    |

| l0y1    | A-219     | K-132     | G-131     | D-90      | AAAA   |
| l0y1    | A-219     | K-132     | G-131     | D-90      | BBBB   |
| l0y1    | A-219     | K-132     | G-131     | D-90      | CCCC   |
| l0y1    | A-219     | K-132     | G-131     | D-90      | DDDD   |
| * 1nu8:  | A-59      | K-68      | G-65      | D-56      | AAAA   |
| * 1nu8:  | A-59      | K-68      | G-65      | D-56      | BBBB   |
| 1sg9:    | G-248     | K-202     | G-225     | D-250     | BBBB   |
| 1to0:G   | G-60      | K-56      | G-112     | D-57      | BBBB   |
| 1to0:G   | G-60      | K-56      | G-112     | D-57      | DDDD   |
| 1vhq:    | G-60      | K-56      | G-112     | D-57      | GGFGG  |
| 1vim:    | A-118     | K-112     | G-87      | D-114     | CCCC   |
| * 1xtl:  | A-169     | K-53      | G-52      | D-171     | BBBB   |

Table 5 (part 1 of 7)
|          | CPM       | TSM       |
|----------|-----------|-----------|
| IF159    | V-71      | E-94      | I-86      | B-113     | AAAA     |
| 11xR     | V-34      | E-55      | I-50      | E-48      | BBBB     |
| 1awv     | V-97      | D-58      | I-90      | E-85      | AAAA     |
| 1awv     | V-97      | D-58      | I-90      | E-85      | BBBB     |
| 1ntz     | V-51      | E-19      | I-13      | D-10      | AAAA     |
| 1ntz     | V-610     | E-313     | I-310     | BBBB      |
| 1ntz     | V-651     | E-613     | D-610     | CCCC      |
| 1oly     | V-134     | E-175     | I-213     | E-209     | AAAA     |
| 1pmj     | V-26      | E-29      | D-30      | BBBB      |
| 1ntz     | V-33      | E-13      | I-7       | E-36      | AAAA     |
| 1ntz     | V-33      | E-13      | I-7       | E-36      | BBBB     |
| 1ntz     | V-34      | E-55      | I-50      | E-48      | AAAA     |
| 1ntz     | V-610     | E-313     | I-310     | BBBB      |
| 1ntz     | V-651     | E-613     | D-610     | CCCC      |
| 1oly     | V-134     | E-175     | I-213     | E-209     | AAAA     |
| 1pmj     | V-26      | E-29      | D-30      | BBBB      |
| 1ntz     | V-33      | E-13      | I-7       | E-36      | AAAA     |
| 1ntz     | V-33      | E-13      | I-7       | E-36      | BBBB     |
| 1ntz     | V-34      | E-55      | I-50      | E-48      | AAAA     |
| 1ntz     | V-610     | E-313     | I-310     | BBBB      |
| 1ntz     | V-651     | E-613     | D-610     | CCCC      |
| 1oly     | V-134     | E-175     | I-213     | E-209     | AAAA     |
| 1pmj     | V-26      | E-29      | D-30      | BBBB      |

Table 5 (part 2 of 7)
|    | CPM       | TSM       |
|----|-----------|-----------|
|    | lvGo:A     | 2.3854    |
|    | lvUp:A     | 2.3854    |
|    | lmqm:A     | 14.1443   |
|    | lmqm:D     | 14.3092   |
|    | ljsm:A     | 15.1853   |
|    | ljso:A     | 15.3448   |
|    | ljso:A     | 52.4113   |
|    | ljsm:A     | 52.5309   |
|    | lmqm:A     | 52.8660   |
|    | lmqm:D     | 52.8783   |
|    | lvGo:A     | 92.1400   |
|    | lvUp:A     | 92.2602   |
|lvka:A | 4.4976 | 1.152:B | 16.4414 |
|liuk:A | 5.1423 | 1.17d:A | 16.4568 |
|lsqh:A | 5.5636 | 1.174:A | 16.6065 |
|lf89:A | 7.2532 | 1.lvdhE | 16.7980 |
|lf89:B | 7.2532 | 1.lqy1:A | 19.4749 |
|lmaz:B | 7.7333 | 1.lwae:B | 20.4061 |
|lywf:A | 7.8641 | 1.lwae:A | 20.6489 |
|lmaz:B | 8.4137 | 1.lz94:F | 21.6438 |
|ly62:B | 8.6742 | 1.tc5:B | 24.7067 |
|linj:B | 9.9203 | 1.tc5:C | 24.9828 |
|lreo2:A | 10.2528 | 1.tc5:B | 25.1026 |
|lreo2:A | 10.7170 | 1.tc5:A | 25.1029 |
|lreo2:B | 11.5752 | lv2a:B | 25.7388 |
|lreo2:B | 12.1043 | ly7p:C | 30.3085 |
|lreo2:B | 12.1841 | ly7p:A | 31.4373 |
|lreo2:B | 12.6492 | ly7p:B | 35.9452 |
|lreo2:C | 13.8226 | ly7p7:A | 36.1629 |
|lreo2:C | 14.9818 | ly7p7:B | 36.6625 |
|lreo2:C | 15.5316 | ly7p7:A | 36.7537 |
|lreo2:D | 15.5665 | ly7p7:B | 36.7843 |
|lreo2:A | 16.1886 | ly7p7:A | 36.9103 |
|lreo2:B | 16.1575 | ly7p7:B | 36.9277 |

* lvka:A | ly7p:A | 1.2232 | linj:B | 18.3348 |
* liuk:A | ly7p:C | 1.3810 | linj:B | 20.5706 |
* lsqh:A | lpt8:B | 1.7413 | lte5:B | 20.7348 |
* lf89:A | lpt8:A | 1.7680 | lts8:A | 23.2399 |
* lf89:B | lpt5:A | 1.7724 | lte5:A | 23.3779 |
* lmaz:B | lpt7:A | 1.7724 | luc2:B | 27.4541 |
* lywf:A | lpt7:B | 1.7724 | luc2:B | 27.8532 |
* lmaz:A | lpt5:B | 1.8035 | lfs9:B | 28.6851 |
* ly62:B | ltc5:A | 2.3320 | lfs9:A | 28.8255 |
* linj:B | ltc5:B | 2.3940 | lj74:A | 31.5884 |
* lreo2:A | ltc5:C | 2.5465 | lmaz:A | 31.8142 |
* lreo2:A | ltc5:D | 2.6225 | lj74:A | 32.0144 |
* lreo2:B | ltc5:D | 2.6225 | lj74:A | 32.0144 |
* lreo2:B | ltc5:D | 2.6225 | lj74:A | 32.0144 |

\[25\]
| Protein | α1-Sequence | β1-Sequence | κ1-Sequence | σ1-Sequence | (cont’d.) |
|---------|-------------|-------------|-------------|-------------|----------|
| 1f89    | R-226       | Y-235       | W-204       | G-236       | AAAA     |
| 1f89    | R-526       | Y-535       | W-504       | G-536       | BBBB     |
| * 1link | R-135       | Y-35        | M-127       | V-128       | AAAA     |
| 1j74    | R-61        | Y-63        | M-126       | G-52        | AAAA     |
| 1j7d    | R-61        | Y-63        | M-126       | G-52        | AAAA     |
| 1mzg    | Q-49        | S-47        | W-55        | V-55        | AAAA     |
| 1mzg    | Q-49        | S-47        | W-55        | V-55        | BBBB     |
| 1nxj    | R-130       | S-40        | W-90        | V-54        | BBBB     |
| 1o62    | Q-39        | S-40        | M-242       | V-41        | AAAA     |
| 1gy1    | R-17        | Y-207       | W-184       | V-21        | AAAA     |
| 1lsfs   | Q-12        | S-10        | W-32        | V-14        | AAAA     |
| * 1sqh  | O-99        | S-97        | W-98        | G-80        | AAAA     |
| 1oc62   | Q-1080      | S-1079      | M-1027      | G-1077      | AAAA     |
| 1tc5    | Q-99        | Y-41        | M-39        | V-40        | AAAA     |
| 1tc5    | O-99        | Y-41        | M-39        | V-40        | BBBB     |
| 1tc5    | Q-99        | Y-41        | M-39        | V-40        | COCC     |
| 1tc5    | Q-99        | Y-41        | M-39        | V-40        | DDDD     |
| 1te5    | Q-239       | Y-191       | W-247       | V-245       | BBBB     |
| 1uc2    | R-27        | Y-29        | M-63        | V-62        | AAAA     |
| 1uc2    | R-27        | Y-29        | M-63        | V-62        | BBBB     |
| 1udh    | R-52        | Y-62        | W-51        | V-40        | DDDD     |
| 1udh    | R-52        | Y-62        | W-51        | V-40        | BBBB     |
| * 1yka  | R-84        | S-55        | M-48        | G-81        | AAAA     |
| 1w4k    | R-80        | S-82        | W-128       | V-127       | COCC     |
| 1wue    | R-1235      | S-1259      | W-1200      | V-1267      | AAAA     |
| 1wue    | R-2235      | S-2259      | W-2288      | V-2287      | BBBB     |
| 1wuf    | O-1088      | S-1271      | M-1270      | G-1269      | AAAA     |
| 1wuf    | Q-2088      | S-2271      | M-2270      | G-2269      | BBBB     |
| * 1y6z  | R-72        | Y-74        | W-5         | V-99        | BBBB     |
| 1y9v    | Q-219       | S-216       | M-189       | G-217       | AAAA     |
| 1y9v    | R-235       | Y-238       | M-126       | G-230       | AAAA     |
| 1z94    | Q-136       | S-133       | W-130       | G-129       | PPPPP     |
| 2ax1    | R-77        | Y-80        | M-106       | V-107       | AAAA     |

Table 5 (part 3 of 7)
| CPM | TSM |
|-----|-----|
| 2acl:A | 0.0000 | lxsl:A | 22.4551 |
| 2acl:C | 0.0000 | lxsl:A | 22.5071 |
| 2acl:E | 0.0000 | lxsl:B | 22.5071 |
| 2acl:G | 0.0000 | lxsl:B | 22.5071 |
| 1fib:Y | 19.7966 | 1fib:Y | 23.3751 |
| 1lxk:A | 22.0000 | 1lxk:A | 23.5462 |
| 1lxk:E | 22.0000 | 1lxk:A | 25.8555 |
| 1lyf:A | 22.9565 | 1lyf:A | 25.3136 |

Table 5 (part 4 of 7)
# Heme-Bound Nitric Oxide-Binding Proteins

| CPM  | TSM  |
|------|------|
| 1zo1:B | 19.6326 |
| 1ozw:A | 20.3294 |
| 1zo1:A | 20.6529 |
| 1xk3:A | 20.9409 |
| 1xbr:A | 2.2055 |
| 1xbr:B | 2.2095 |
| 1xbr:A | 5.3211 |
| 1xbr:B | 5.3922 |
| 1xbx:A | 5.3955 |
| 1xbx:B | 5.4021 |
| 1xbr:A | 5.4154 |
| 1xbr:B | 5.5215 |
| 1iku:A | 6.4279 |
| 1iku:A | 6.5303 |
| 1qy9:A | 7.2523 |
| 2a67:D | 7.3117 |
| 2a67:D | 7.4473 |
| 2a67:A | 7.4492 |
| 1qy9:B | 7.4775 |
| 1qy9:A | 8.0550 |
| 1qy9:A | 9.3680 |
| 1qy9:A | 10.3653 |
| 1qy9:A | 10.4986 |
| 1zbs:A | 10.5114 |
| 1qy9:C | 10.8514 |
| 1qy9:A | 10.9059 |
| 1lw7:A | 10.9503 |
| 1qyv:B | 10.9545 |
| 1xewe:A | 11.0664 |
| 1qyv:D | 11.2290 |
| 2arz:A | 1.7076 |
| 1xew:A | 1.7920 |
| 2arz:B | 2.1786 |
| 1xew:B | 4.5930 |
| 2brw:A | 5.9513 |
| 1lan:A | 9.4262 |
| 1lan:B | 9.5238 |
| 1zsw:A | 10.2977 |
| 1xe8:A | 11.5042 |
| 1xel:A | 17.0775 |
| 1xel:B | 17.0775 |
| 1xel:D | 17.0775 |
| 1xel:C | 17.1655 |
| 1xew:A | 17.6712 |
| 1xew:B | 18.1097 |
| 1xew:B | 18.2039 |
| 2a67:D | 19.3501 |
| 2a67:A | 19.4883 |
| 1u9c:A | 25.9022 |
| 1qvy:D | 28.3240 |
| 1qvy:A | 28.8285 |
| 1qvy:A | 29.2536 |
| 1qvy:B | 29.3370 |
| 1qvy:C | 29.3023 |
| 1zbr:A | 68.8762 |
### heme-NO-binding proteins (cont’d.)

| Code | heme NO | C-met | F-met | H-met | BB | BB | BB |
|------|---------|-------|-------|-------|----|----|----|
| 1o6l | G-350   | C-101 | F-80  | H-336 | BB | BB | BB |
| 1o6h | G-350   | C-101 | F-80  | H-336 | BB | BB | BB |
| 1qvf | G-25    | H-108 | P-17  | Y-13  | AA | AA | AA |
| 1qvf | G-25    | H-108 | P-17  | Y-13  | BB | BB | BB |
| 1qvf | G-25    | H-108 | P-17  | Y-13  | CC | CC | CC |
| 1qvf | G-25    | H-108 | P-17  | Y-13  | DD | DD | DD |
| 1qvr | G-25    | H-108 | P-17  | Y-13  | BB | BB | BB |
| 1qvr | G-25    | H-108 | P-17  | Y-13  | BB | BB | BB |
| 1qvr | G-25    | H-108 | P-17  | Y-13  | BB | BB | BB |
| 1qvr | G-25    | H-108 | P-17  | Y-13  | BB | BB | BB |
| 1qvc | G-22    | H-96  | P-100 | H-12  | AA | AA | AA |
| 1uan | G-20    | C-18  | F-213 | R-198 | AA | AA | AA |
| 1uan | G-20    | C-18  | F-213 | R-198 | BB | BB | BB |
| 1vkh | G-112   | H-36  | P-50  | H-109 | BB | BB | BB |
| 1wdc | G-89    | H-432 | P-87  | R-63  | AA | AA | AA |
| lxcC | H-39    | C-73  | P-71  | H-79  | AA | AA | AA |
| lxcC | H-39    | C-73  | P-71  | H-79  | BB | BB | BB |
| lxcC | H-39    | C-73  | P-71  | H-79  | CC | CC | CC |
| lxcC | H-39    | C-73  | P-71  | H-79  | DD | DD | DD |
| 1x8e | G-93    | H-202 | P-178 | H-42  | AA | AA | AA |
| 1x6b | G-34    | H-120 | P-110 | H-13  | AA | AA | AA |
| 1x6b | G-34    | H-13  | P-118 | H-120 | BB | BB | BB |
| 1xtl | H-120   | H-86  | P-114 | H-121 | AA | AA | AA |
| 1xtl | H-120   | H-86  | P-114 | H-121 | BB | BB | BB |
| 1xtl | H-120   | H-86  | P-114 | H-121 | CC | CC | CC |
| 1xtl | H-120   | H-86  | P-114 | H-121 | DD | DD | DD |
| 1xtm | G-123   | H-112 | P-114 | H-120 | AA | AA | AA |
| 1zbs | G-100   | C-106 | P-125 | Y-277 | AA | AA | AA |
| 1szw | H-46    | H-9   | F-256 | R-253 | AA | AA | AA |
| 2arz | H-201   | C-40  | P-110 | Y-17  | AA | AA | AA |
| 2arz | H-201   | C-40  | P-110 | Y-17  | BB | BB | BB |
| 2hd4 | G-191   | C-205 | P-148 | R-136 | AA | AA | AA |
| 2f9c | G-212   | H-213 | P-191 | R-215 | AA | AA | AA |
| 2f9c | G-212   | H-213 | P-191 | R-215 | DD | DD | DD |

Table 5 (part 5 of 7)
## Unbound Nitric Oxide-binding proteins

| CPM       | TSM       |
|-----------|-----------|
| 1zgn:B    | 25.5828   | 1zgn:A    | 5.7669   |
| 1zgn:A    | 25.2577   | 1zgn:B    | 5.0896   |
| 1xa0:A    | 1.9106    | * 1uc2:A  | 10.9601  |
| 1xa0:B    | 2.0976    | * 1vim:B  | 13.3793  |
| 1tzz:A    | 5.3056    | * 1tzz:B  | 15.7609  |
| 1tzz:B    | 6.4536    | * 1zn6:A  | 16.0147  |
| 1wu8:A    | 7.2081    | 1tzz:A    | 17.1025  |
| 1vim:B    | 13.9373   | 1vbk:A    | 19.7980  |
| * 1uc2:A  | 14.7380   | 1vbk:B    | 20.7845  |
| * 1zn6:A  | 15.0978   | 1wu8:A    | 62.0812  |
| * 1vbk:B  | 20.0751   | 1xa0:B    | 93.1935  |
| * 1vbk:A  | 21.0505   | 1xa0:A    | 94.8195  |

| 1tzz | Y-2061 | R-2060 | I-2016 | G-2059 | BBBB |
|------|--------|--------|--------|--------|------|
| * 1uc2 | Y-451 | R-408  | I-74   | G-407  | AAAA |
| 1vbk | Y-8   | R-50   | I-116  | G-49   | AAAA |
| 1vbk | Y-8   | R-50   | I-116  | G-49   | BBBB |
| 1vim | Y-47  | R-44   | I-63   | G-46   | BBBB |
| * 1zn6 | Y-156 | R-61   | I-153  | G-60   | AAAA |

Table 5 (part 6 of 7)
<Ligand>-binding protein family

|        | CPM                      | TSM                      |
|--------|--------------------------|--------------------------|
| training structures w/ their CPI | ▲▲▲▲ ▲▲ | ▲▲▲▲ ▲▲ |
| application structures that tested positive for the LBS and matched the CPI of a training structure | ▲▲▲▲ ▲▲ ▲▲ | ▲▲▲▲ ▲▲ ▲▲ |

Training structures (in reverse order) w/ their TSi

Application structures that tested positive for the LBS and matched the TSi of a training structure

Application structures that tested positive for the LBS and matched both the CPI and TSi of a training structure

Four amino acids in the 3D SM

Protein chain

Table 5 (part 7 of 7)
| PDS ID | LBS Detected | PDS Header | Source Organism | Remarks | Published Reference |
|--------|--------------|------------|----------------|---------|---------------------|
| 1RU8   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 11-DEC-94 | Pyrococcus furiosus | PUTATIVE ATP PYRONEPHOSPHATASE, NOS TARGET | none |
| 1XTL3  | GTP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 23-OCT-04 | Bacillus subtilis | P194 MUTANT OF HYPOTHETICAL SUPEROXIDE DISMUTASE | none |
| 1FL9   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 13-AUG-00 | Haemophilus influenzae | THE BVE PROTEIN, HYPOTHETICAL PROTEIN | Proteins. 2002 Aug 14(2):220-6 |
| 1MW6   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 01-OCT-02 | Haemophilus influenzae | STRUCTURE OF THE HYPOTHETICAL PROTEIN | none |
| 1NF2   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 12-OCT-02 | Thermotoga maritima | PROTEIN TNO201, A PHOSPHATASE WITH A NEW FOLD | none |
| 101Y   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 13-FEB-03 | Thermotoga maritima | CONSERVED HYPOTHETICAL PROTEIN TM175, A PUTATIVE GLUTAMINE AMINO TRANSFERASE | Proteins. 2004 Mar 154(4):601-5 |
| 1RKL   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 24-MAY-03 | Pyrococcus aerophilum | PROTEIN PAE5A2 WITH THREE DISULFIDE BONDS | Proteins. 2005 Sep 8(1-2):300-6 |
| 1RKO   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 23-MAY-03 | Escherichia coli | PROTEIN ECA113, A HETERO-ODDIN FOLD | none |
| 1T57   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 24-MAY-03 | Methanococcus thermautotrophicus | CONSERVED PROTEIN MVTH75 | none |
| 1TQ6   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 12-OCT-02 | Bacteroides H277RV | HYPOTHETICAL PROTEIN RY599, NOS | none |
| 1UM6   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 04-JUL-04 | Thermus thermophilus H28 | PROTEIN TT0010, A PHENYLACETIC ACID DEGRADATION PROTEIN | J Mol Biol. 2005 Sep 9(1):213-23 |
| 2CVL   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 06-JUN-05 | Thermus thermophilus H8 | PROTEIN TT0017, A TRANSLATION INITIATION INHIBITOR | none |
| 1UK    | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 05-MAY-02 | Thermoanaerobacterium | HYPOTHETICAL PROTEIN TT0088, A CONSERVED COA-BINDING PROTEIN | none |
| 1SOH   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 05-APR-04 | Desulfovibrio desulfuricans | HYPOTHETICAL PROTEIN CO44514P (CO44514P), NOS | none |
| 1VKA   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 05-MAY-02 | Homo sapiens | HYPOTHETICAL PROTEIN J00617, N TERMINAL FRAGMENT, FRAGMENT-ASSOCIATED PROTEIN | none |
| 1Y6Z   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 07-DEC-04 | Pseudomonas putida | C-TERMINAL DOMAIN OF PUTATIVE HET SHock PROTEIN | none |
| 1NHJ   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 31-DEC-02 | Bacillus subtilis | THE YOLK PROTEIN | none |
| 1NX4   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 08-FEB-03 | Enorma carotovora | CIRABEREN SYNTYHASE (CIRI) W JELLY ROLL | J Biol Chem. 2003 Jun 27(23):20238-45 |
| 1NX8   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 12-FEB-03 | Enorma carotovora | CIRABEREN SYNTYHASE (CIRI) W JELLY ROLL | J Biol Chem. 2003 Jun 27(23):20238-45 |
| 1RLH   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 05-APR-04 | Thermoplasma acidophilum | A CONSERVED HYPOTHETICAL PROTEIN FROM | none |
| 1RV9   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 12-MAY-02 | Nocardioides aquisolvens | A CONSERVED HYPOTHETICAL PROTEIN NAB0700, W ALPHA-BETA-ALPHA STRUCTURE | none |
| 1U11   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 24-JUN-04 | Pseudomonas aeruginosa | HYPOTHETICAL PROTEIN | none |
| 1U6L   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 30-JUN-04 | Pseudomonas aeruginosa | HYPOTHETICAL PROTEIN | none |
| 1IVM   | ATP          | STRUCTURAL GENOMICS, UNKNOWN FUNCTION 01-OCT-02 | Archaeoglobus fulgidus | HYPOTHETICAL PROTEIN | J Neurochem. 2003 Dec 25(3):1471-80 |
| SM  | Type        | Genus            | Species | Function       | Structure | Protein                        | Description                                                                 |
|-----|-------------|------------------|---------|----------------|-----------|--------------------------------|----------------------------------------------------------------------------|
| 1WU8| retinoic ac  | *N. gonorrhoea*   |         |                |           |                                | Hypothetical protein, 3D structure unknown                                  |
| 1YBT| retinoic ac  | *Bacteroides*     |         |                |           |                                | Hypothetical protein, 3D structure unknown                                  |
| 1YBE| retinoic ac  | *Xanthomonas*     |         |                |           |                                | Hypothetical protein, 3D structure unknown                                  |
| 1ZRM| retinoic ac  | *Enterococcus*    |         |                |           |                                | Hypothetical protein, 3D structure unknown                                  |
| 2EVR| retinoic ac  | *Nostoc*          |         |                |           |                                | Hypothetical protein, 3D structure unknown                                  |
| 1UAN| heme-NO     | *Thermus*         |         |                |           |                                | Hypothetical protein, 3D structure unknown                                  |
| 1VKH| heme-NO     | *Thermus*         |         |                |           |                                | Hypothetical protein, 3D structure unknown                                  |
| 1WDT| heme-NO     | *Thermus*         |         |                |           |                                | Hypothetical protein, 3D structure unknown                                  |
| 1XE8| heme-NO     | *Thermus*         |         |                |           |                                | Hypothetical protein, 3D structure unknown                                  |
| 1XQB| heme-NO     | *Thermus*         |         |                |           |                                | Hypothetical protein, 3D structure unknown                                  |
| 1XTL| heme-NO     | *Thermus*         |         |                |           |                                | Hypothetical protein, 3D structure unknown                                  |
| 1ZSY| heme-NO     | *Thermus*         |         |                |           |                                | Hypothetical protein, 3D structure unknown                                  |
| 2ARZ| heme-NO     | *Thermus*         |         |                |           |                                | Hypothetical protein, 3D structure unknown                                  |
| 2B4V| heme-NO     | *Thermus*         |         |                |           |                                | Hypothetical protein, 3D structure unknown                                  |
| 1UC2| unbound NO  | *Pseudomonas*     |         |                |           |                                | Hypothetical protein, 3D structure unknown                                  |
| 1ZR6| unbound NO  | *Bordetella*      |         |                |           |                                | Hypothetical protein, 3D structure unknown                                  |

* A heme-NO (retinoic acid) 3D model has also been detected in this structure; see 1XTL* below.
* A 3D SM corresponding to monomer 1 in the interleukin-2 homodimer complex (2NKB) was also detected in this structure (see Reyes, V.M., 2008c).
* A 3D SM corresponding to monomer 1 in the RAP-RAF-RAS-RBD complex (RAP-RasRBD complex, with RAS-RAF-RAS binding domain, 1C1Y) was also detected in this structure (see Reyes, V.M., 2008c).
* The specific GTP LBS detected is the 3D SM corresponding to that in the small, Ras-type G-protein family.
* The specific ATP LBS detected is the 3D SM corresponding to that in the ser/th protein kinase family.

Table 6.