Identification of Novel Small Molecule Inhibitors of Core-Binding Factor Dimerization by Computational Screening against NMR Molecular Ensembles

Ryan H. Lilien[*,†,‡] **Mohini Sridharan**[*,**] **Bruce R. Donald**[*,‡,§,¶,‖]

March 9, 2004

Abstract

The long development process of novel pharmaceutical compounds begins with the identification of a lead inhibitor compound. Computational screening to identify those ligands, or small molecules, most likely to inhibit a target protein may benefit the pharmaceutical development process by reducing the time required to identify a lead compound. Typically, computational ligand screening utilizes high-resolution structural models of both the protein and ligand to fit or 'dock' each member of a ligand database into the binding site of the protein. Ligands are then ranked by the number and quality of interactions formed in the predicted protein-ligand complex. It is currently believed that proteins in solution do not assume a single rigid conformation but instead tend to move through a small region of conformation space [14]. Therefore, docking ligands against a static snapshot of protein structure has predictive limitations because it ignores the inherent flexibility of the protein [14]. A challenge, therefore, has been the development of docking algorithms capable of modeling protein flexibility while balancing computational feasibility. In this paper, we present our initial development and work on a molecular ensemble-based algorithm to model protein flexibility for protein-ligand binding prediction. First, a molecular ensemble is generated from molecular structures satisfying experimentally-measured NMR constraints. Second, traditional protein-ligand docking is performed on each member of the protein's molecular ensemble. This step generates lists of ligands predicted to bind to each individual member of the ensemble. Finally, lists of top predicted binders are consolidated to identify those ligands predicted to bind multiple members of the protein's molecular ensemble. We applied our algorithm to identify inhibitors of Core Binding Factor (CBF) among a subset of approximately 70,000 ligands of the Available Chemicals Directory. Our 26 top-predicted binding ligands are currently being tested experimentally in the wetlab by both NMR-binding experiments (^{15}N-edited Heteronuclear Single-Quantum Coherence (HSQC)) and Electrophoretic Gel Mobility Shift Assays (EMSA). Preliminary results indicate that of approximately 26 ligands tested, three induce perturbations in the protein's NMR chemical shifts indicative of ligand binding and one ligand (2-amino-5-cyano-4-tertbutyl thiazole) causes a band pattern in the EMSA indicating the disruption of CBF dimerization.

Dartmouth Computer Science Department Technical Report No: TR2004-492
`http://www.cs.dartmouth.edu/reports/reports.html`

[*]Dartmouth Computer Science Department, Hanover, NH 03755, USA.
[†]Dartmouth Medical School, Hanover, NH 03755, USA.
[‡]Dartmouth Center for Structural Biology and Computational Chemistry, Hanover, NH 03755, USA.
[§]Dartmouth Department of Chemistry, Hanover, NH 03755, USA.
[¶]Dartmouth Department of Biological Sciences, Hanover, NH 03755, USA.
[‖]Corresponding author, `Bruce.R.Donald@dartmouth.edu`. This work is supported by grants to B.R.D. from the National Institutes of Health (GM-65982), and the National Science Foundation (IIS-9906790, EIA-0102710, EIA-0102712, EIA-9818299, EIA-9802068, and EIA-0305444).
[**]Author's current address: University of California San Francisco Department of Biophysics, San Francisco, CA 94143, USA.

1 Introduction

Core Binding Factor is a heterodimeric transcription factor involved in hematopoesis and is composed of α and β subunits (Figure 1). The CBF-α:CBF-β heterodimer complex binds DNA, affecting cell growth through regulation of homeobox (HOX) genes [21]. Translocations in the CBF-α and CBF-β genes are frequently oncogenic and are implicated in several subtypes of leukemia [20, 21, 28, 9]. In particular, Acute Myelomonocytic Leukemia (AMML) is associated with the oncogenic gene fusion CBF-β-MYH11 formed through a chromosome 16 inversion [20]. In the corresponding fused protein, the smooth muscle myosin heavy chain (SMMHC) protein encoded by the MYH11 gene is covalently attached to CBF-β. While the CBF-α binding site of the fusion protein remains functionally intact, the CBF-α:CBF-β-SMMHC complex does not properly regulate gene expression. Immunofluorescence localization experiments [15, 1] demonstrate that the CBF-β domain of the CBF-β-SMMHC fusion protein oligomerizes with wild-type α-subunits while the SMMHC domain binds actin filaments in the cell's cytoplasm. These interactions cause the entire complex (including CBF-α) to be sequestered outside the nucleus thereby vitiating transcription [15, 1]. Therefore the ultimate goal of this computer-assisted drug design effort was to design a small-molecule inhibitor to disrupt the complex formed by the wild-type CBF-α with the oncoprotein CBF-β-SMMHC. Disruption of the complex would allow copies of CBF-α not bound to CBF-β-SMMHC to associate with DNA. Alone, CBF-α is capable of regulating gene expression albeit to a lesser extent than the CBF-α:CBF-β complex [6, 30].

One challenge faced in the design of a small molecule to disrupt CBF-α:CBF-β dimerization was that at the time we began this project, small molecule dimerization inhibitors were essentially unknown. The X-ray crystallographer Gregory Petsko wrote: "To my knowledge there is no small-molecule drug that has yet been designed to disrupt a protein-protein interaction. Such targets will be of increasing importance as our understanding of signal transduction and transcriptional regulation deepens" [24]. As the field of drug-design progresses, it will transition to include modification of cell signaling and transcription factors such as CBF. Our work on CBF represents a first step in this direction.

The use of molecular ensembles for drug design presents a challenge. Traditional drug design algorithms spend significant effort removing conformations from consideration so as to reduce the computational time required for ligand binding evaluation. In contrast, ensemble-based approaches seek to accurately sample conformation space by including as many molecular conformations as possible. In this work, we represent protein flexibility by an ensemble of low-energy molecular conformations generated from solution *Nuclear Magnetic Resonance* (NMR) spectroscopy. Therefore each conformation of the molecular ensemble is consistent with the experimental NMR restraints and represents a protein conformation possibly assumed in solution. By including a modest number of NMR-consistent conformations we incorporate a degree of molecular flexibility without suffering a combinatorial explosion.

This work presents a description of our use of molecular ensembles to model a target protein's flexibility in a novel Computer-Aided Drug Design (CADD) approach. In this paper we describe the *in silico* design of a ligand to disrupt dimerization of the wild-type CBF-α and CBF-β subunits. Our approach was successfully used to discover a novel inhibitor of CBF dimerization. Our inhibitor could be useful in itself, in that one could potentially disrupt the healthy transcription factor with a small ligand, allowing new *in vivo* studies of AMML. Our inhibitor could also serve as a lead compound to inhibit the oncogenic form of CBF-β. In Section 2 we present previous work on

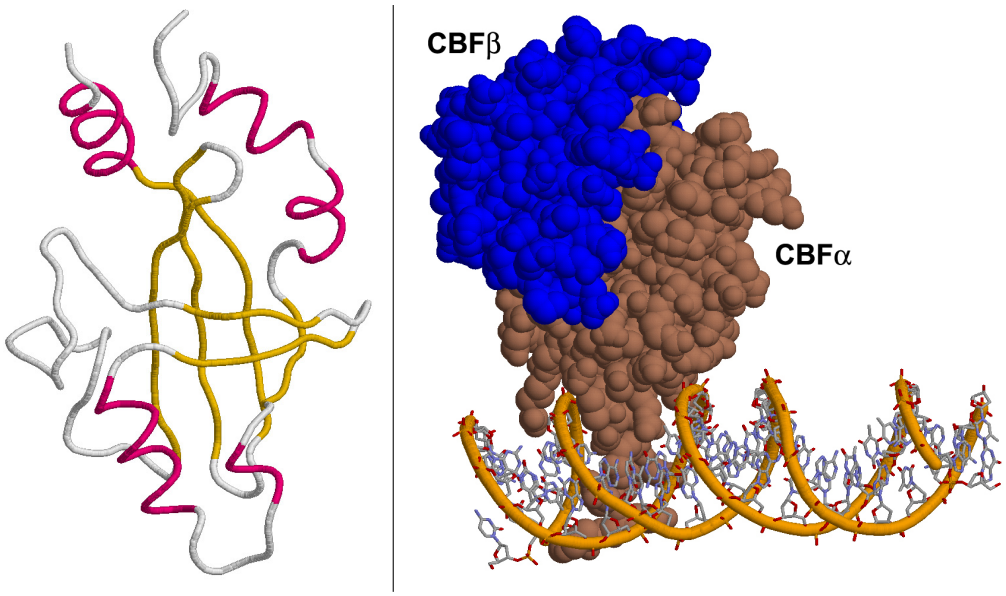


Figure 1: **Core Binding Factor.** (Left) Backbone trace of the first model of the NMR solution structure of CBF-β (PDB: 2JHB [13]). (Right) The CBF complex (CBF-α (brown) and CBF-β (blue)) are shown in complex with DNA (PDB: 1IO4 [29]). While CBF-α makes intimate contact with the DNA, CBF-β 'piggybacks' on CBF-α and does not directly interact with the DNA. The binding of CBF-β to CBF-α increases the binding affinity of CBF-α for DNA [6, 30].

molecular modeling using conformational ensembles. Our ensemble scoring approach is described in Section 3. Section 3.1 presents our method for ensemble generation, Section 3.2 and 3.3 provide details on our use of the LUDI search algorithm, the ACD database, and our method of identifying top ensemble binders, Section 3.4 describes the two wetlab experiments (SAR by NMR and EMSA) performed. Finally, in Section 4 we present the results of applying our ensemble search strategy to the identification of dimerization inhibitors for CBF.

2 Previous Work

The intuition behind the use of ensembles is straightforward: a ligand predicted to bind multiple structures of an ensemble has the potential to bind the protein as it moves through a region of conformation space. We can biophysically ground the use of ensembles with the Ergodic hypothesis. The Ergodic hypothesis states that phenomena, such as the binding energy ΔG, are experimentally measured as a time average and that the time average is equal to the ensemble average. Therefore, in principle, accurately computing the ensemble average allows one to accurately predict experimentally observed phenomena.

The ensemble average specified in the Ergodic hypothesis is weighted by the Boltzmann probability of observing each state. That is for some measured phenomenon, G,

$$G_{obs} = \frac{1}{\tau}\int_{t_0}^{t_0+\tau} G(t)dt = \int_c P(c)G_c,$$

where the first integral represents the time average over a timestep τ and $G(t)$ is the value of G

at time t. The second integral is taken over all of conformation space, $P(c)$ is the probability of observing state c, and G_c is the value of G associated with state c. In a Boltzmann distribution, the probability $P(c)$ is $\exp(-E_c/RT)/Z$ where E_c is the energy of conformation c, R is the gas constant, T is the temperature in Kelvin, and Z is the partition function, $Z = \int_c \exp(-E_c/RT)$. Thus by the Ergodic hypothesis, the value of a phenomenon measured as a time average of one molecule is equal to the value measured as a Boltzmann-weighted ensemble average over a sufficiently large set of molecules. The challenge that arises when attempting to compute a complete ensemble average is computing the integral over all of conformation space. Thus attempts to sample conformational space at 'high-yield' positions have been developed [19, 17]. These high-yield conformations are those with high probability of occurrence (i.e., those with low energies). Therefore there are two steps to an ensemble based scoring scheme. First, one needs to develop a method for sampling conformation space (generating the ensemble) and second, one needs to devise a way of combining or utilizing those conformations (scoring the ensemble).

2.1 Ensemble Generation

There are three main methods for generating ensembles to sample conformational space. The first technique for generating a molecular ensemble is to perform a Molecular Dynamics (MD) simulation starting from a known or hypothesized conformation and to save snapshots of the system throughout the MD simulation [25]. This technique can generate an arbitrary number of molecular conformations. The drawbacks of MD-based ensembles is that there is no guarantee they evenly or completely sample conformation space. That is, the conformations sampled are inherently biased based on the starting conformation. As a result, MD-generated ensembles sample a relatively local region of conformation space. The second method of generating molecular ensembles uses amino acid rotamers. Modeling side-chain motions with a discrete set of side-chain conformations, or a rotamer library, allows a reduced representation of the amino-acid conformational space. In a rotamer library, each amino acid is represented by a set of commonly assumed conformations [26, 22]. These conformations are typically mined from analysis of high resolution protein structures in the protein databank; alternatively rotamers may be generated from a sampling of low energy conformations as scored with an energy function.

In the final ensemble generating method, an ensemble can be created from one or more NMR experiments or multiple X-ray crystallography experiments[18, 25]. A published NMR 'structure' is an ensemble of structures each of which similarly satisfies the experimental constraints. Flexible regions of a protein generally do not allow sufficient geometric constraints to be measured by NMR and therefore show increased variance in position among conformations in an NMR ensemble. As an example, both the N- and C-termini typically have fewer NMR constraints, correspondingly most published NMR ensembles show increased positional variance at the N- and C-termini. The structures of an NMR- or crystallography-generated ensemble may demonstrate local structural differences but tend not to capture extremely large motions. As an ensemble-generating technique, the use of multiple NMR or crystallographic structures has a limitation in that a relatively small number of conformations may be sampled (typically on the order of 20); however, these techniques have the advantage that the crystallographic structures represent protein conformations assumed in the crystal and the NMR structures are consistent with experimentally measured conformational constraints.

2.2 Ensemble Scoring

Several techniques have been developed for utilizing molecular ensembles. When the work presented in this report was performed (1998) a very limited corpus of previous work existed on ensemble scoring [16]. However, after the completion of our research, a number of approaches, similar to our method, have been published [5, 16, 8, 23]. These approaches perform conventional docking against each member of an ensemble and typically compute the average or best interaction energy between the protein and ligand. Ensembles have also been used to construct a dynamic pharmacophore and screen a ligand database [7].

The ensemble model presented in this report represents an initial simple ensemble model, and has been replaced in our lab by our more recently developed ensemble scoring method [19, 17]. The more sophisticated method, called K^*, is derived to be an approximation to the true association (binding) constant K_A by expressing each species' chemical potential as a function of the species' partition function and solving for the equilibrium condition [19]:

$$K^* = \frac{q_{PL}}{q_P q_L}$$

$$q_{PL} = \sum_{b \in B} \exp(-E_b/RT), \quad q_P = \sum_{f \in F} \exp(-E_f/RT), \quad q_L = \sum_{l \in L} \exp(-E_l/RT), \tag{1}$$

where B, F, and L represent rotamer-based ensembles for the bound protein-ligand complex (PL), the free protein (P), and the free ligand (L) respectively, E_s is the energy of conformation s, R is the gas constant, and T is the temperature in Kelvin. The accuracy with which K^* approximates K_A is proportional to the accuracy of the partition function approximation used. We have also developed an efficient deterministic approximation algorithm, capable of approximating K^* to arbitrary precision [19]. In [19], K^* was used to model molecular flexibility to predict protein-ligand binding for protein redesign. K^* identified two mutation sequences of the Gramicidin Synthetase A Phenylalanine Adenylation domain with altered ligand specificity; the ligand specificity of the redesigned proteins switched from phenylalanine to leucine. Predicted mutation sequences were created in the wetlab and experimentally exhibited the desired change in substrate specificity.

3 Methods

The overall ligand screening algorithm is shown in Figure 2. First, a suite of NMR experiments are performed to obtain sequence-specific resonance assignments and extract geometrical constraints (e.g., NOEs and scalar couplings) on the structure of a target protein. Second, the molecular dynamics simulated annealing program DYANA [11] generates an ensemble of twenty molecular conformations which best satisfy the NMR experimental constraints. Third, a single-structure database screening algorithm, LUDI [4], is used to screen the *Available Chemicals Directory* (ACD) database (MDL Information Systems, San Leandro, CA) against each of the twenty DYANA generated conformations. Fourth, the top ACD ligands (as scored by LUDI) are consolidated to find ligands which rank among the top LUDI-predicted binders for a majority of DYANA-generated CBF-β conformations. Finally, these top ligands were purchased and screened using two wetlab experimental techniques: *Structure Activity Relationship by NMR* (SAR by NMR) [27, 12] and *Electrophoretic Gel Mobility Shift Assay* (EMSA).

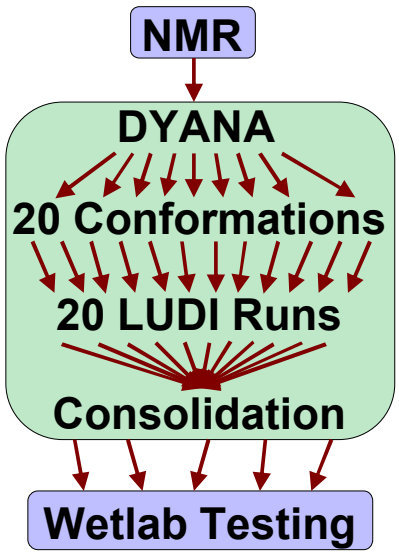


Figure 2: **Ensemble-Based CADD Algorithm.** The computational steps developed and described in this paper are shown in green while the wetlab steps performed by our collaborators in the Speck and Bushweller labs are shown in blue. Geometric constraints on the protein's conformation are measured by NMR spectroscopy, these constraints are utilized by the molecular dynamics program DYANA [11] to generate an ensemble of 20 protein conformations. The ligand docking program LUDI is run on each member of the ensemble and generates lists of ligands predicted to bind each conformation. Ligands which are predicted to bind multiple protein conformations are identified during the consolidation step. These ligands are then tested experimentally in the wetlab.

3.1 Ensemble Generation

The structure of CBF-β was solved in 1999, by two labs including our collaborators in John Bushweller's lab in the Department of Chemistry at Dartmouth [13, 10]. We therefore had access to both the initial and final solution NMR ensembles for CBF-β [13]. The first (initial) ensemble was generated from preliminary NMR data whereas the second (final) ensemble was computed from the final NMR data and is the ensemble deposited in the protein databank (PDB: 2JHB [13]). In this paper, both the initial and final molecular ensembles were used in CADD screening; each ensemble contained 20 low-energy conformations of CBF-β. NMR spectra were collected by the Bushweller lab using a 500 MHz UnityPlus Varian spectrometer.

Conceptually the molecular structures in each ensemble are the CBF-β structural models that best satisfy the NMR constraints calculated by experimentally-measured solution NMR. The computer program DYANA [11] generates molecular structures using molecular dynamics subject to the experimentally measured NMR constraints. While all 20 DYANA-generated conformations satisfy the NMR constraints to a similar extent, these conformations vary slightly in geometry. For comparison, three conformations of the first NMR ensemble are shown in Figure 3. These conformations illustrate that while the overall protein fold is conserved across generated conformations, portions of the protein have fewer constraints and are therefore modeled in multiple positions. Specifically, the loop on the left side of each structure in Figure 3 assumes three drastically different conformations among the three structures. The residues of the CBF-α binding site (Section 3.2 below) are shown

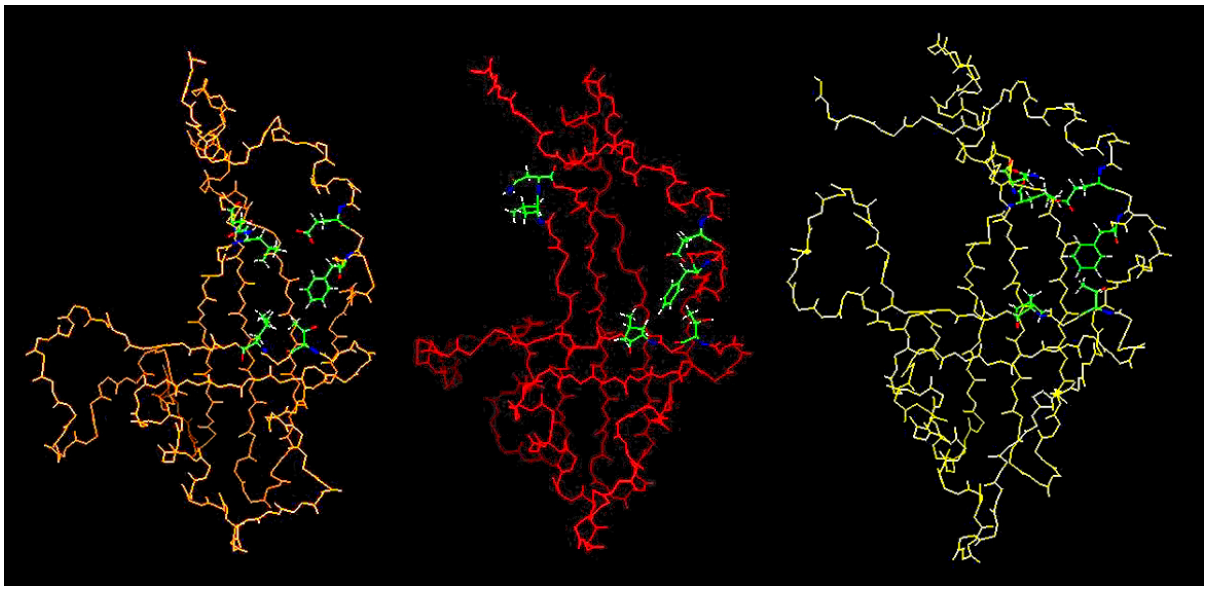


Figure 3: **Conformations of Core Binding Factor Beta.** Three conformations of CBF-β generated by DYANA using experimentally measured NMR constraints. All three structures have a similar orientation, significant flexibility is seen in the loop on the left side of the protein and in the residues (green) of the CBF-α binding site.

in green in Figure 3; these residues also assume different conformations in each structure. The conformational differences seen in the CBF-α binding site are exactly the structural differences we want to exploit during the CADD search.

3.2 LUDI Search of the Available Chemicals Directory (ACD)

At the time this research was performed, the hypothesized dimerization interface (CBF-α binding site) had been localized only by analysis of site-directed mutants [31]. Residues in the vicinity of Leu66 and Asn106 (2JHB numbering) were identified as important to binding in that when the amino-acid type of these residues was altered, the binding of CBF-α to CBF-β was affected [31]. With this knowledge, we used LUDI [2, 4] to search the ACD database for ligands predicted to bind in the dimerization interface of each member of the NMR-based ensemble. The version of LUDI implemented as part of the InsightII software package (Accelrys, San Diego, CA) was used in our search. This implementation of the LUDI algorithm consists of two stages. First, LUDI identifies *interaction sites* among atoms of the protein and ligand. Interaction sites are labeled as: lipophilic-aliphatic, lipophilic-aromatic, hydrogen-bond donor, or hydrogen-bond acceptor. Second, interaction sites are matched so that, for example, a hydrogen-bond donor is matched with a hydrogen-bond acceptor. LUDI matches interaction sites by computing the distance between three ligand interaction sites and then searching for three complementary interaction sites in the protein which share a similar geometry [2]. Matching is done geometrically thereby avoiding the need to minimize energy functions. Positioned ligands are then scored based on the number and types of interaction sites satisfied as well as their deviation from ideal geometries. Additional detail of the LUDI algorithm and scoring function can be found in [2, 3, 4].

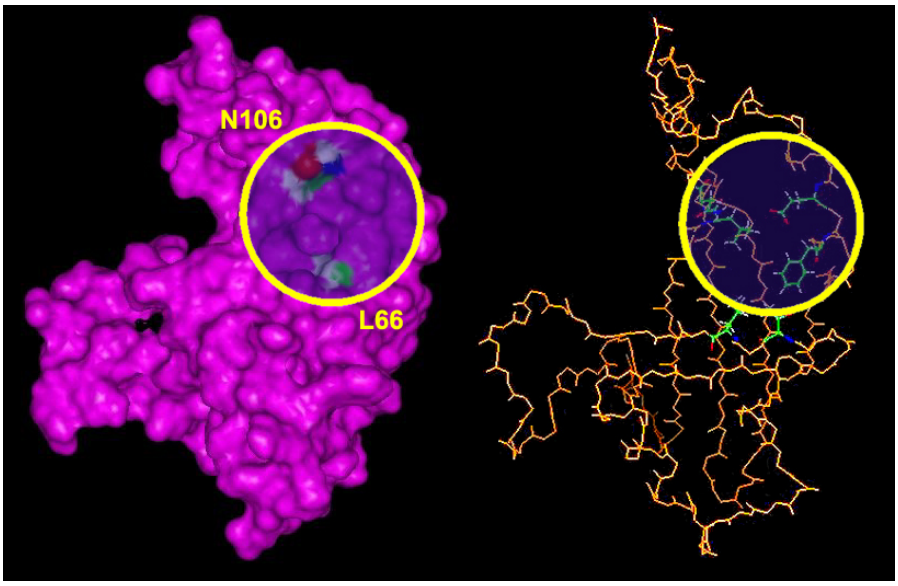


Figure 4: **LUDI's Ligand Search Sphere for CBF-β.** CBF-β is shown with the ligand search sphere circled in yellow. (Left) Solvent accessible surface of CBF-β with residues L66 and N106 colored and the LUDI search sphere circled in yellow. (Right) A second view containing a backbone trace of CBF-β with ligand search sphere circled in yellow.

The InsightII distribution of LUDI searches a subset of the ACD database containing approximately 70,000 small organic ligands with two or fewer rotatable bonds (for reference, the complete ACD database contains approximately 240,000 commercially available chemicals). To direct the search and limit the region of the protein to which a ligand may bind, LUDI requires that the user specify a *ligand search sphere*. With the goal of disrupting the dimerization of CBF-α:CBF-β, LUDI was instructed to search for ligands capable of binding CBF-β in the hypothesized CBF-α binding site. All runs used a ligand search sphere of radius of 8Å, the maximum allowed, centered midway between residues Leu66 and Asn106 (Figure 4). This region includes the residues identified by analysis of site-directed mutants as being important in CBF-α:CBF-β dimerization.

Subsequent to the completion of our computational screening, both chemical shift perturbation experiments[1] on CBF-β [13] and solution of the CBF-α:CBF-β crystal structure (Figure 5) to 3.00Å (PDB: 1IO4 [29]) confirmed that the region enclosed by the ligand search sphere is indeed included in the CBF-α:CBF-β dimerization interface.

3.3 Consolidation

The output of each LUDI run is a list of ligands with predicted binding affinities above a cutoff threshold. Therefore, running LUDI on all twenty members of the molecular ensemble generates twenty sets of ligands each containing molecules predicted to bind CBF-β with high affinity. A computer program was written to analyze these result files. Because we were looking for ligands

[1]Chemical shift perturbation experiments identify the regions of the protein involved in complex formation by examining changes in NMR chemical shifts upon complex formation. A more detailed description is provided in Section 3.4.

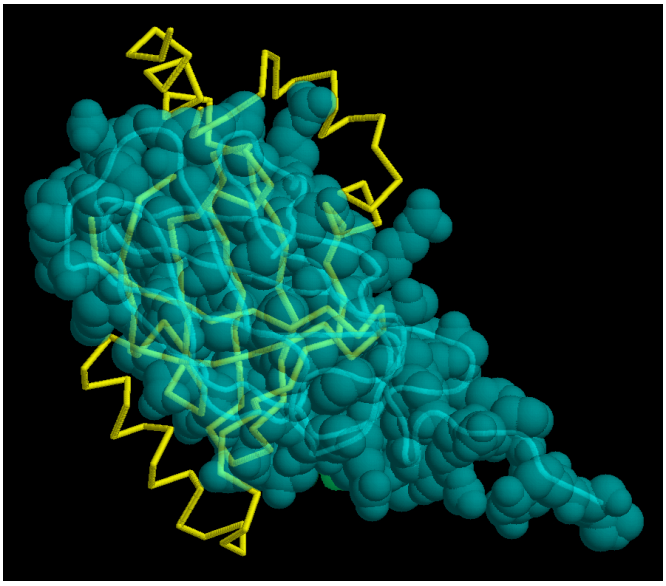


Figure 5: **Core Binding Factor.** The bound conformation of CBF-α:CBF-β is shown (PDB: 1IO4). The backbone of CBF-β is shown in yellow and is positioned in an orientation similar to Figures 3 and 4. CBF-α is shown in transparent cyan spacefill with the backbone trace in light-cyan. The ligand search sphere used in each LUDI search (Figure 4) is included in the dimerization interface region of 1IO4.

predicted to bind multiple members of the ensemble, for each ligand that appears in any of the twenty lists, the number of runs which contain each ligand is computed as well as the average LUDI score among those runs. The complete list of ligands is then sorted first on the number of lists in which each ligand appears and second on each ligand's average LUDI score in those runs. The result of this analysis is an ordering of ligands (from best to worst) which may then be tested in the wetlab.

3.4 Wetlab Testing

SAR by NMR. The SAR by NMR assay measures the ability of one molecule (e.g., a ligand) to perturb the chemical shifts of a second molecule (e.g., the target protein) [27, 12]. The phenomenon measured in a SAR by NMR assay is therefore similar to that of chemical shift perturbation experiments. Both experiments exploit the fact that the local electronic environment plays a dominant role in affecting the chemical shift measured for each atom. The binding of CBF-α (or a ligand) tends to change the local electronic environment of the residues involved in the dimerization interface while leaving unchanged the local electronic environment of those residues not involved in binding. NMR spectroscopy is thus used to measure chemical shifts for each residue both before and after binding (i.e., two experiments are performed). By looking for atoms with chemical shifts that differ significantly between the bound (holo) and unbound (apo) experiments one can identify the residues involved in binding[2].

[2]A simplified version of chemical perturbation analysis and SAR by NMR is described here. In practice, some residues not involved in binding may also experience a change in chemical shift due to overall changes in protein conformation.

Those ligands that do not bind the protein do not alter the local electronic environment and the chemical shifts of the protein should appear the same as they did before addition of the ligand. The magnitude of the change in chemical shift as well as the magnitude of the peaks in the original and moved positions can provide a rough estimate of the strength of ligand binding. Most important, in addition to answering the question of *if* the protein binds a target molecule, chemical shift perturbation analysis and SAR by NMR also answers the question of *where* the binding occurs. Therefore, the primary use of the SAR by NMR experiment by our collaborators in this research was to screen the computationally identified ligands for binding to CBF-β.

The typical NMR experiment performed in chemical perturbation analysis and SAR by NMR is the ^{15}N-edited Heteronuclear Single-Quantum Coherence (HSQC). This experiment correlates ^{15}N atoms with bound hydrogens and can be recorded in less than an hour. After collecting data for solutions of pure protein and protein plus ligand, the easily-identified backbone amide (^{15}N,H) peaks are examined for changes in chemical shift. In the backbone amide region of the HSQC spectrum, each residue has a single corresponding peak. When performing these assays, only the protein should be ^{15}N labeled; the ligand should remain unlabeled. This means that no special processing of the ligand is required[3].

Electrophoretic Gel Mobility Shift Assay. The EMSA measures the ability of a protein to bind DNA (Figure 6). The fundamental principle of the EMSA is that when DNA is loaded into an agarose gel and an electrical potential is applied, the negatively charged DNA will move towards the positive electrode with a speed inversely proportional to the mass of the DNA complex (i.e., larger DNA complexes move more slowly through the gel). Therefore after a fixed period of time, unbound DNA will move the furthest through the gel whereas DNA bound to one or more proteins will move more slowly. In the case of CBF, lanes are run with fixed concentrations of DNA, CBF-α, and ligand, but with varying concentrations of CBF-β. Ligands which bind CBF-β and prevent the binding of CBF-α to CBF-β reduce the effective concentration of free CBF-β requiring increased concentrations of CBF-β to achieve the same amount of CBF-β:CBF-α:DNA complex formation. Therefore a tightly-binding ligand will defer the formation of the CBF-β:CBF-α:DNA complex (the slow moving band) until high concentrations of CBF-β are present. With increasing concentration of CBF-β the CBF-β:CBF-α:DNA complex is formed and a slow moving band appears on the gel.

4 Results

The ensemble search process (Sections 3.1 to 3.3) was performed twice, once using an ensemble generated from preliminary NMR data and a second time using an ensemble generated from the final NMR data. For each of the twenty CBF-β structures in each ensemble, LUDI was used to search the ACD database as provided with the InsightII software distribution. The top binders identified by each run were then consolidated to identify the ligands which best bind the ensemble (Section 3.3). Each LUDI run required approximately 5 hours of wall-clock time when run on a single processor SGI O2 (MIPS R10000 175MHz Processor). The computation of the consolidation step requires only seconds to complete.

[3]When screening a large database of ligands by SAR by NMR, a binary search strategy can be employed [12]. In this strategy, multiple ligands are tested simultaneously in the same NMR experiment. If a positive result is observed (i.e., chemical shifts move between the protein only and protein + ligand experiments) then the pool of ligands may be split into two sets and each screened again. If an experiment for a set of ligands shows no chemical shift change then that entire set of ligands can be removed from consideration. In our screening this binary search strategy was not employed.

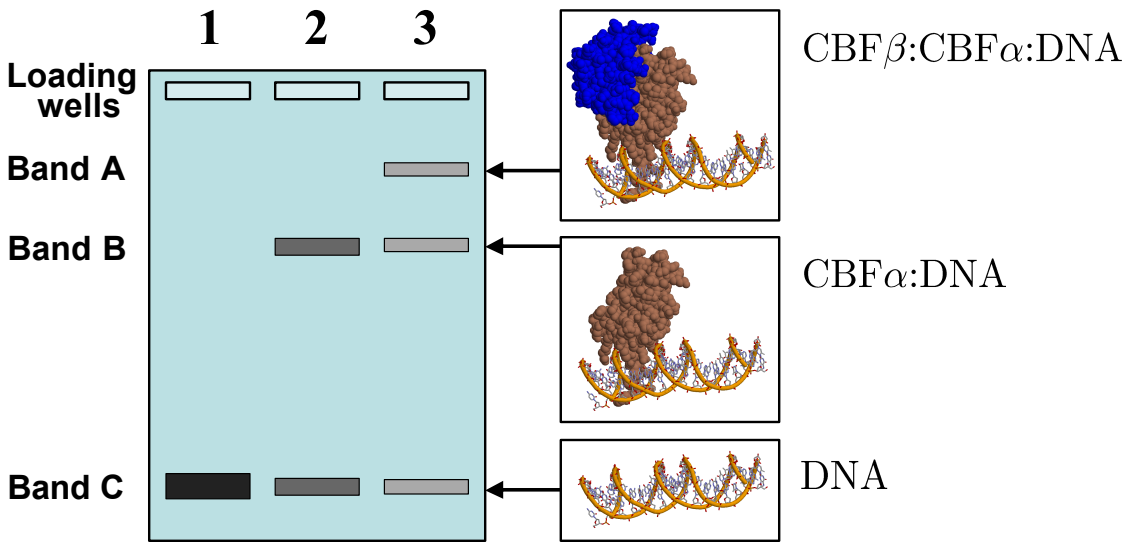


Figure 6: **Electrophoretic Gel Mobility Shift Assay (EMSA).** A cartoon of an EMSA. Three samples (1, 2, 3) are shown as having been run in three different lanes on the gel. In this example, lane 1 contains DNA only, lane 2 contains CBF-α and DNA, and lane 3 contains CBF-α, CBF-β, and DNA. When an electric potential is applied across the gel, the negatively-charged DNA is attracted to the positive electrode. The speed with which the DNA moves through the gel (in this diagram starting from the top and moving toward the bottom of the gel) is dependent upon the size of the DNA complex. When the DNA is free (band C) it moves the farthest. When CBF-α is bound to the DNA, the complex is not able to move as far (band B). When DNA is bound to the CBF-α:CBF-β complex the entire complex moves the slowest (band A). Therefore, in the absence of a dimerization inhibitor a lane with CBF-α, CBF-β, and DNA will have a strong band A. If a ligand disrupts the binding of CBF-α and CBF-β then one should see a more intense band B and a lighter band A. Because binding is never complete bands with partial complexes are often seen (i.e., band C in lane 2 and bands B and C in lane 3).

A peculiarity in the way LUDI returns results forced us to adopt the following search strategy. LUDI will output a maximum of only 940 ligand hits. Unfortunately, LUDI does not output the *top* 940 hits but rather the *first* 940 hits at which point the search terminates. Therefore, in order to identify the top hits for each member of the ensemble, the LUDI search parameter `Min_Score` required adjustment[4]. LUDI considers ligands with scores larger (better) than `Min_Score` as hits and writes these molecules to an output file. Setting the value of `Min_Score` too high results in no ligands being classified as hits whereas setting the value of `Min_Score` too low results in too many hits. Therefore, `Min_Score` must be set low enough that some ligands are classified as hits yet high enough that the LUDI ACD search does not prematurely terminate on account of 940 compounds having already been found. Searches were performed with a `Min_Score` of 350, those searches which returned fewer than 20 hits were rerun with a lower `Min_Score` threshold, down to a minimum `Min_Score` of 270 (Table 1).

Ligands which appeared in multiple LUDI hit lists were identified as described in Section 3.3.

[4]All LUDI searches were performed with the default value for each search parameter except the `Min_Score` parameter.

PDB	Min_Score	Num. Hits
001_NQ	340	43
002_NQ	340	120
003_NQ	320	71
004_NQ	340	56
005_NQ	340	63
006_NQ	310	40
007_NQ	340	117
008_NQ	330	85
009_NQ	270	46
010_NQ	320	118
011_NQ	320	190
012_NQ	290	60
013_NQ	320	32
014_NQ	320	136
015_NQ	270	43
016_NQ	320	167
017_NQ	320	139
018_NQ	320	61
019_NQ	270	14
020_NQ	280	47

Table 1: **LUDI Runs.** The twenty LUDI runs performed on the initial ensemble. The 'PDB' column lists the identifier of each ensemble member, the '`Min_Score`' column lists the value of the `Min_Score` LUDI parameter used, and the 'Num. Hits' column lists the number of ligands which satisfy the `Min_Score` LUDI threshold.

All ligands identified as hits for at least eight of the twenty members of the ensemble are shown in Tables 2 and 3. While structural differences between the two ensembles prevent the two lists of top binders from being identical, six ligands, **3**, **5**, **6**, **7**, **8**, and **11** do appear in both lists. Therefore, 26 unique ligands were identified through the two runs.

All but two of the 26 ligands have a diamide motif, capable of forming two hydrogen bonds (Figures 7 and 8). The predicted binding modes of these ligands are similar, the rigid ring structure sterically fits into a small concave binding pocket while the diamide motif forms two hydrogen bonds with Glu17. By wedging into the CBF-α binding site, these ligands have the potential to disrupt native CBF-α-CBF-β contacts thereby inhibiting dimerization.

At this point, the list of top predicted binders was handed off to our biological collaborators in the labs of Nancy Speck and John Bushweller. Wetlab testing of the 26 identified leads is still in progress. Thus far most of the 26 compounds (Tables 2 and 3) have been purchased and tested by SAR by NMR in John Bushweller's lab. Preliminary SAR by NMR results indicated that of the 26 tested ligands, three induce changes in chemical shifts for atoms of residues in the proposed binding site. These results are indicative of binding in the CBF-α:CBF-β dimerization site. These three compounds were next tested for their ability to disrupt CBF dimerization using an EMSA. These experiments were performed by Yen-Yee Tang in Nancy Speck's lab. One of the three tested ligands, **21** (2-amino-5-cyano-4-tertbutyl thiazole) (Table 3 and Figure 8) was found to inhibit CBF dimerization at millimolar concentrations. Because 2-amino-5-cyano-4-tertbutyl thiazole both binds CBF-β in the dimerization interface region and inhibits dimer formation it can serve as a lead compound for pharmaceutical development. The next stage of development will be to increase the binding strength and specificity of 2-amino-5-cyano-4-tertbutyl thiazole. Thus a

Top Predicted Binders from Ensemble 1

Molecule Number	MFCD ID	Avg. Score (Num Found)	Molecule Description
1	MFCD00142855	387.30 (10)	2-AMINO-1-METHYLBENZIMIDAZOLE
2	MFCD00101498	385.10 (10)	MAYBRIDGE NRB 01318
3	MFCD00024137	376.60 (10)	1-AMINOISOQUINOLINE
4	MFCD00159978	369.00 (10)	SPECS CIF6564
5	MFCD00005789	383.67 (9)	2-AMINO-6-METHYLBENZOTHIAZOLE
6	MFCD00130077	370.78 (9)	MAYBRIDGE BTB 11174
7	MFCD00185989	404.75 (8)	2-AMINONAPHTHO(2,1-D)THIAZOLE
8	MFCD00193713	399.62 (8)	N-PHENYL-BENZAMIDINE
9	MFCD00205354	385.50 (8)	MAYBRIDGE BTB 12069
10	MFCD00183946	376.62 (8)	6-ISOPROPYL-BENZOTHIAZOL-2-YLAMINE
11	MFCD00006322	371.00 (8)	2-AMINO-4,6-DIMETHYLPYRIDINE
12	MFCD00129026	367.75 (8)	2-AMINO-4-PROPYL PYRIDINE
13	MFCD00086324	363.00 (8)	5-AMINO-2,3-DIMETHYLINDOLE
14	MFCD00274654	348.00 (8)	2,2',3,3',5,5'-HEXAMETHYL-(1,1'-BIPHENYL)-4,4'-DIOL

Top Predicted Binders from Ensemble 2

Molecule Number	MFCD ID	Avg. Score (Num Found)	Molecule Description
15	MFCD00224194	345.79 (14)	2-AMINO-6-METHYL-4-PHENYL-NICOTINONITRILE
8	MFCD00193713	344.07 (14)	N-PHENYL-BENZAMIDINE
16	MFCD00224193	344.08 (13)	2-AMINO-4-PHENYL-6,7-DIHYDRO-5H-(1)PYRINDINE-3-CARBONITRILE
17	MFCD00233438	346.83 (12)	SPECS CIF3563
18	MFCD00004054	351.73 (11)	1-(2-NAPHTHYL)-2-THIOUREA
19	MFCD00006460	347.64 (11)	3-AMINO-5,6-DIMETHYL-1,2,4-TRIAZINE
3	MFCD00024137	342.18 (11)	1-AMINOISOQUINOLINE
11	MFCD00006322	341.36 (11)	2-AMINO-4,6-DIMETHYLPYRIDINE
6	MFCD00130077	336.64 (11)	MAYBRIDGE BTB 11174
7	MFCD00185989	324.36 (11)	2-AMINONAPHTHO(2,1-D)THIAZOLE
5	MFCD00005789	324.50 (10)	2-AMINO-6-METHYLBENZOTHIAZOLE
20	MFCD00136481	361.44 (9)	1-(1-(2-NAPHTHYL)ETHYLIDENE)-3-THIOSEMICARBAZIDE
21	MFCD00214700	340.22 (9)	2-AMINO-5-CYANO-4-TERTBUTYL THIAZOLE
22	MFCD00142856	329.00 (9)	2-AMINO-3-METHYLNAPHTHO(1,2)-IMIDAZOLE
23	MFCD00244146	353.50 (8)	2-AMINO-6-TERT-BUTYL-3-CYANO-4-PHENYLPYRIDINE
24	MFCD00039680	334.88 (8)	2-AMINO-4-PHENYLTHIAZOLE
25	MFCD00051730	329.25 (8)	MAYBRIDGE RJC 00685
26	MFCD00168154	291.75 (8)	4,4'-(1,3-ADAMANTANEDIYL)DIPHENOL)

Table 2: **Top Predicted Binders.** These tables list all ligands predicted to bind at least eight of the twenty members of each molecular ensemble. Column 'Molecule Number' is a reference molecule number used in the figures and text of this paper, Column 'MFCD ID' is the MFCD identifier used to search the ACD database for the molecular name, Column 'Avg. Score (Num Found)' displays the fragment's average LUDI score and the number of CBF-β structures the ligand was predicted to bind, Column 'Molecule Description' lists the molecular name or unique catalog identifier. The molecules **3**, **5**, **6**, **7**, **8**, and **11** appear as top predicted binders for both ensembles, therefore the two searches produced a total of 26 unique ligands. Compound **21** (2-amino-5-cyano-4-tertbutyl thiazole) (Figure 8) was found to bind CBF-β by SAR by NMR and to inhibit CBF dimerization by EMSA.

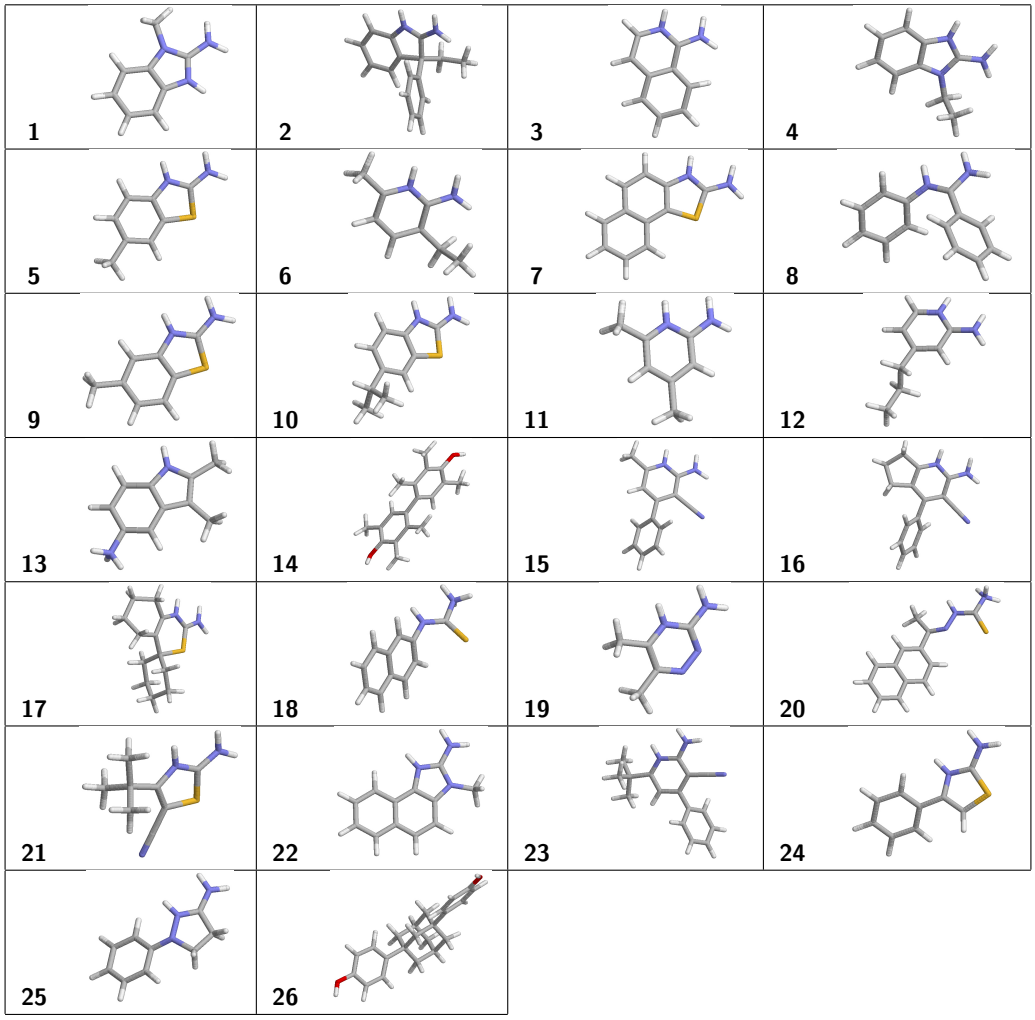
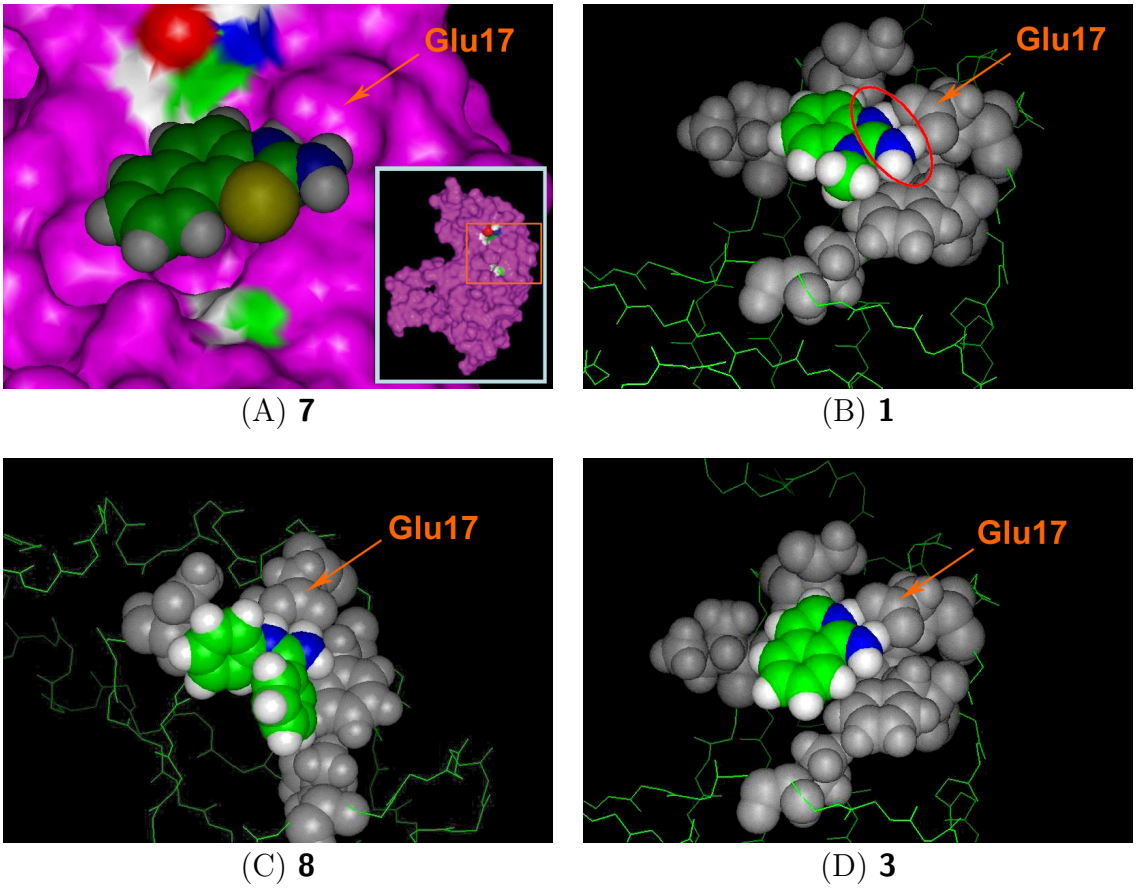


Table 3: **Molecular Structures of Top Predicted Binders.** The 26 ligands which are predicted to bind at least eight of twenty structures in the molecular ensemble. Most of these ligands possess a relatively rigid aliphatic ring structure with an amide motif capable of forming two hydrogen bonds. Molecule numbers correspond to the 'Molecule Number' column in Table 2.



(A) **7** (B) **1**

(C) **8** (D) **3**

Figure 7: **Predicted Binding Modes of Top Ligands.** The predicted binding modes for ligands (A) **7**, (B) **1**, (C) **8**, and (D) **3**. The ligand is shown in color (C: green, N: blue, H: white, S: yellow). In (A) **7** is shown with the molecular surface of the protein (pink with residues 66 and 106 colored). The region of the protein shown is highlighted by an orange box in the inset figure. In (B), (C), and (D), the CBF-β backbone is shown in green with the residues of the binding pocket in grey spacefill. All four molecules have a diamide motif (circled in panel B) that forms hydrogen bonds with Glu17; this binding mode is predicted to be present in all 24 ligands possessing a diamide motif.

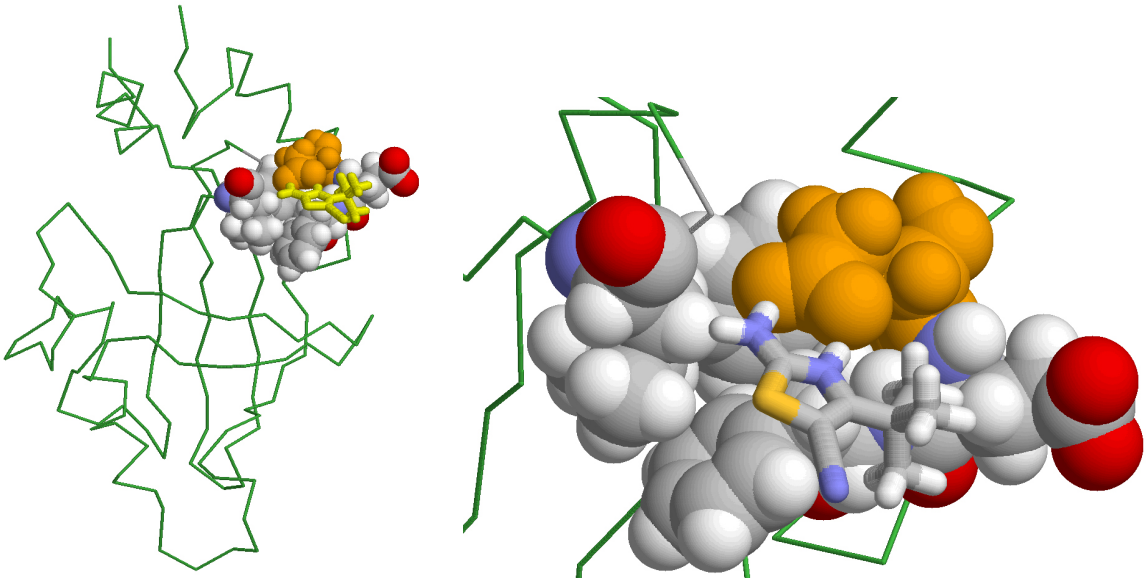


Figure 8: **Predicted Binding of 21.** The predicted binding mode for ligand **21** (2-amino-5-cyano-4-tertbutyl thiazole). The location of the binding mode is shown in (A); **21** is shown in yellow wireframe while the binding site residues are shown in spacefill with Glu17 in orange. Panel (B) shows a closeup of the binding site in the same orientation; the binding site residues are once again shown in spacefill (Glu17 is orange) and **21** is shown in wireframe with standard CPK colors. The diamide motif of **21** is predicted to form two hydrogen bonds with Glu17. Molecule **21** was shown to bind CBF-β by SAR by NMR, and was shown to disrupt CBF-α:CBF-β dimerization by EMSA.

lead compound has been identified.

5 Conclusion

The molecular ensemble docking method presented here was successfully applied to identify three lead compounds that bind CBF-β, one of which disrupts CBF dimerization as desired. The use of an NMR-based ensemble allowed for the modeling of molecular flexibility while keeping the algorithm's runtime manageable. Our results — a lead compound capable of binding and disrupting the wild-type protein-protein interface — were validated in the wet-lab using electrophoretic gel mobility shift assays and SAR by NMR (^{15}N-HSQC chemical shift perturbation).

The ensemble-based scoring method presented in this report is quite simple. Nonetheless, this approach successfully identified three ligands capable of binding the target protein as measured by SAR by NMR with an excellent hit rate of approximately 1 in 9. The simple model presented here provided the groundwork upon which our more advanced, biophysically derived, K^* ensemble model [19] was developed.

In this report we have shown that ligands predicted to bind multiple members of an NMR ensemble have a high experimental *in vitro* hit rate. It therefore would be interesting to investigate the binding affinity of ligands predicted to tightly bind only one or two ensemble members and compare their actual *in vitro* binding to those ligands predicted to bind a large number of ensemble members (i.e., those ligands identified in this work). Such a control experiment could provide additional support for the use of NMR-based molecular ensembles.

We note that the ensemble-based strategy presented here, screening ligands against each member of a low-energy NMR molecular ensemble and then searching for ligands predicted to bind well among multiple members of the ensemble, does not critically depend on the use of LUDI. While LUDI worked very well for us in this set of experiments, in theory, as ligand docking algorithms advance, the LUDI screening step could be replaced with any more recently developed scoring algorithm. The overall screening method would then proceed as follows: 1) generate a molecular ensemble using NMR spectroscopy, 2) perform docking studies using any desired docking algorithm on each ensemble member, 3) consolidate the top predicted results to identify the best ensemble binders, 4) screen top predicted binders *in vitro*.

6 Acknowledgments

We thank John Bushweller and Nancy Speck for their collaboration and support of wetlab experiments for the determination of the CBF-β structure and ligand binding assays. We thank Xuemei Huang for her work in the Bushweller lab solving the 2JHB CBF-β structure, and Yen-Yee Tang in the Speck lab for performing the EMSA experiments. We also thank Amy Anderson and Chris Bailey-Kellogg for helpful discussions and comments on drafts.

References

[1] N. Adya, T. Stacy, N. Speck, and P. Liu. The leukemic protein core binding factor β (CBFβ)-smooth-muscle myosin heavy chain sequesters CBFα2 into cytoskeletal filaments and aggregates. *Mol. Cell. Biol.*, 18:7432–7443, 1998.

[2] H.J. Böhm. The computer program LUDI: A new method for the de novo design of enzyme inhibitors. *J. Comput. Aided Mol. Des.*, 6:61–78, 1992.

[3] H.J. Böhm. LUDI: Rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput. Aided Mol. Des.*, 6:593–606, 1992.

[4] H.J. Böhm. On the use of LUDI to search the fine chemicals directory for ligands of proteins of known three-dimensional structure. *J. Comput. Aided Mol. Des.*, 8:623–632, 1994.

[5] D. Bouzida, P. Rejto, S. Arthurs, A. Colson, S. Freer, D. Gehlhaar, V. Larson, B. Luty, P. Rose, and G. Verkhivker. Computer simulations of ligand-protein binding with ensembles of protein conformations: A monte carlo study of HIV-1 protease binding energy landscapes. *Int. J. Quantum Chem.*, 72:73–84, 1999.

[6] J. Bushweller. CBF - A biophysical perspective. *Semin. Cell & Dev. Biol.*, 11:377–382, 2000.

[7] H. Carlson, K. Masukawa, and McCammon J. Method for including the dynamic fluctuations of a protein in computer-aided drug design. *J. Phys. Chem. A*, 103:10213–10219, 1999.

[8] H. Claußen, C. Buning, M. Rarey, and T. Lengauer. FlexE: Efficient molecular docking considering protein structure variations. *J. Mol. Biol.*, 308:377–395, 2001.

[9] J. Downing, M. Higuchi, N. Lenny, and A. Yeoh. Alterations of the AML1 transcription factor in human leukemia. *Cell & Dev. Biol.*, 11:347–360, 2000.

[10] M. Goger, V. Gupta, W.Y. Kim, K. Shigesada, Y. Ito, and M. Werner. Molecular insights into PEBP2/CBFβ-SMMHC associated acute leukemia revealed from the structure of PEBP2/CBFβ. *Nature Struct. Biol.*, 6:620–623, 1999.

[11] P. Güntert, C. Mumenthaler, and K. Wüthrich. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.*, 273:283–298, 1997.

[12] P. Hajduk, G. Sheppard, D. Nettesheim, E. Olejniczak, S. Shuker, R. Meadows, D. Steinman, G. Carrera, P. Marcotte, J. Severin, K. Walter, H. Smith, E. Gubbins, R. Simmer, T. Holzman, D. Morgan, S. Davidsen, J. Summers, and S. Fesik. Discovery of potent nonpeptide inhibitors of stromelysin using SAR by NMR. *J. Am. Chem. Soc.*, 119:5818–5827, 1997.

[13] X. Huang, J. Peng, N. Speck, and J. Bushweller. Solution structure of core binding factor β and map of the CBFα binding site. *Nature Struct. Biol.*, 6:624–627, 1999.

[14] W. Jorgensen. Rusting of the lock and key model for protein-ligand binding. *Science*, 254:954–955, 1991.

[15] Y. Kanno, T. Kanno, C. Sakakura, S.C. Bae, and Y. Ito. Cytoplasmic sequestration of the polyomavirus enhancer binding protein 2 (PEBP2)/core binding factor α (CBFα) subunit by the leukemia-related PEBP2/CBFβ-SMMHC fusion protein inhibits PEBP2/CBF-mediated transactivation. *Mol. Cell. Biol.*, 18:4252–4261, 1998.

[16] R. Knegtel, I. Kuntz, and C. Oshiro. Molecular docking to ensembles of protein structures. *J. Mol. Biol*, 266:424–440, 1997.

[17] R. Lilien, A. Anderson, and B.R. Donald. Modeling protein flexibility for structure-based active site redesign. *The 6th Ann. Intl. Conf. on Research in Comput. Mol. Biol. (RECOMB)* in *Currents in Computational Molecular Biology 2002* (ed. L. Florea et al.), pages 122–123, 2002.

[18] R. Lilien, M. Sridharan, J. Huang, J. Bushweller, and B.R. Donald. Computational screening studies for core binding factor beta: Use of multiple conformations to model receptor flexibility. 2000. Poster - 8th International Conference on Intelligent Systems for Molecular Biology ISMB-2000.

[19] R. Lilien, B. Stevens, A. Anderson, and B.R. Donald. A novel ensemble-based scoring and search algorithm for protein redesign, and its application to modify the substrate specificity of the gramicidin synthetase a phenylalanine adenylation enzyme. *Proc. of the 8th Ann. Intl. Conf. on Research in Comput. Mol. Biol. (RECOMB) San Diego, CA, March 27-31*, (in press), 2004.

[20] P. Liu, S. Tarlé, A. Hajra, D. Claxton, P. Marlton, M. Freedman, M. Siciliano, and F. Collins. Fusion between transcription factor CBFβ/PEBP2β and a myosin heavy chain in acute myeloid leukemia. *Science*, 261:1041–1044, 1993.

[21] A. Look. Oncogenic transcription factors in the human acute leukemias. *Science*, 278:1059–1064, 1997.

[22] S. Lovell, J. Word, J. Richardson, and D. Richardson. The penultimate rotamer library. *Proteins*, 40:389–408, 2000.

[23] F. Österberg, G. Morris, M. Sanner, A. Olson, and D. Goodsell. Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in autodock. *Proteins*, 46:34–40, 2002.

[24] G. Petsko. For medicinal purposes. *Nature*, 384 SUPP:7–9, 1996.

[25] M. Philippopoulos and C. Lim. Exploring the dynamic information content of a protein NMR structure: Comparison of a molecular dynamics simulation with the NMR and X-Ray structures of *Escherichia coli* ribonuclease HI. *Proteins*, 36:87–110, 1999.

[26] J. Ponder and F. Richards. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.*, 193:775–791, 1987.

[27] S. Shuker, P. Hajduk, R. Meadows, and S. Fesik. Discovering high-affinity ligands for proteins: SAR by NMR. *Science*, 274:1531–1534, 1996.

[28] N. Speck and T. Stacy. A new transcription factor family associated with human leukemias. *Crit. Rev. Eukar. Gene Expression*, 5:337–364, 1995.

[29] T. Tahirov, T. Inoue-Bungo, H. Morii, A. Fujikawa, M. Sasaki, Kimura K., M. Shiina, K. Sato, T. Kumasaka, M. Yamamoto, S. Ishii, and K. Ogata. Structural analyses of DNA recognition by the AML1/Runx-1 runt domain and its allosteric control by CBFβ. *Cell*, 104:755–767, 2001.

[30] Y. Tang, B. Crute, J. Kelley, X. Huang, J. Tan, J. Shi, K. Hartman, T. Laue, N. Speck, and J. Bushweller. Biophysical characterization of interactions between the core binding factor α and β subunits and DNA. *FEBS Letters*, 470:167–172, 2000.

[31] Y. Tang, J. Shi, L. Zhang, A. Davis, J. Bravo, A. Warren, N. Speck, and J. Bushweller. Energetic and functional contribution of residues in the core binding factor β (CBFβ) subunit to heterodimerization with CBFα. *J. Biol. Chem.*, 275:39579–39588, 2000.