University of Louisville
ThinkIR: The University of Louisville's Institutional Repository

Electronic Theses and Dissertations

8-2010

Determination of in silico rules for predicting small molecule binding behavior to nucleic acids in vitro.

Patrick Andrew Holt
University of Louisville

Follow this and additional works at: https://ir.library.louisville.edu/etd

Recommended Citation

Holt, Patrick Andrew, "Determination of in silico rules for predicting small molecule binding behavior to nucleic acids in vitro." (2010). *Electronic Theses and Dissertations.* Paper 627.
https://doi.org/10.18297/etd/627

This Doctoral Dissertation is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. This title appears here courtesy of the author, who has retained all other copyrights. For more information, please contact thinkir@louisville.edu.

DETERMINATION OF *IN SILICO* RULES FOR PREDICTING SMALL MOLECULE BINDING BEHAVIOR TO NUCLEIC ACIDS *IN VITRO*

By

Patrick Andrew Holt
B.S., The Johns Hopkins University, 2000
M.S., The Johns Hopkins University, 2002
M.S., University of Louisville, 2009

A Dissertation
Submitted to the Faculty of the
Graduate School of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

Department of Biochemistry and Molecular Biology
University of Louisville
Louisville, Kentucky

August 2010

DETERMINATION OF *IN SILICO* RULES FOR PREDICTING SMALL MOLECULE BINDING BEHAVIOR TO NUCLEIC ACIDS *IN VITRO*

By
Patrick Andrew Holt
B.S., The Johns Hopkins University, 2000
M.S., The Johns Hopkins University, 2002
M.S., University of Louisville, 2009

A Dissertation Approved on

May 21, 2010

by the following Dissertation Committee:

Dissertation Director

ACKNOWLEDGMENTS

I am indebted to the collaborative and productive research environment created and fostered by Dr. John O. Trent and Dr. Jonathan B. Chaires which was critical to the success of this work. Few students are fortunate to have such leaders in their field as mentors and provide such extraordinary resources to accomplish their research aims. Their experiences, insights and contributions to all of the work documented here was invaluable. I also thank all of the members of my dissertation committee including Dr. Jonathan B. Chaires, Dr. Thomas E. Geoghegan, Dr. Robert D. Gray, Dr. William L. Dean and Dr. John O. Trent and appreciate their feedback and advice throughout the development of this dissertation. This work would not have been possible without the many scientists in Dr Trent's and Dr. Chaires's labs who gave willingly of their time and expertise. These scientists include: Jonathan M Maguire, Lei Wang, Dr. Nichola C. Garbett, Dr. Luigi Petraccone and Dr. Patricia Ragazzon. I want to thank my wife, Kristen, who has provided so much support and encouragement. I want to thank my parents and brothers who have helped me achieve so many of my goals and dreams. Finally, to my young son Henry: You are the inspiration for everything I do.

ABSTRACT

DETERMINATION OF *IN SILICO* RULES FOR PREDICTING SMALL MOLECULE BINDING BEHAVIOR TO NUCLEIC ACIDS *IN VITRO*

Patrick Andrew Holt

May 21, 2010

The vast knowledge of nucleic acids is evolving and it is now known that DNA can adopt highly complex, heterogeneous structures. Among the most intriguing are the G-quadruplex structures, which are thought to play a pivotal role in cancer pathogenesis. Efforts to find new small molecules for these and other physiologically relevant nucleic acid structures have generally been limited to isolation from natural sources or rationale synthesis of promising lead compounds. However, with the rapid growth in computational power that is increasingly becoming available, virtual screening and computational approaches are quickly becoming a reality in academia and industry as an efficient and economical way to discover new lead compounds. These computational efforts have historically almost entirely focused on proteins as targets and have neglected DNA. We present research here showing that not only can software be utilized for targeting DNA, but that selectivity metrics can be developed to predict the binding mechanism of a small molecule to a DNA target. The software Surflex and Autodock were chosen for evaluation and were demonstrated to be able to accurately reproduce the known crystal structures of several small molecules that bind by the most common nucleic acid interacting mechanisms of groove binding and intercalation. These software were further used to rationalize known affinity and selectivity data of a 67 compound library of compounds for a library of nucleic acid structures including duplex, triplex and quadruplexes. Based upon the known binding behavior of these compounds, *in silico* metrics were developed to classify compounds as either groove binders or intercalators. These rules were subsequently used to identify new triplex and quadruplex binding small molecules by structure and ligand-based virtual screening approaches using a virtual library consisting of millions of commercially available small molecules. The binding behavior of the newly discovered triplex and quadruplex binding compounds was empirically validated using a number of spectroscopic, fluorescent and thermodynamic equilibrium techniques. In total, this research predicted the binding behavior of these test compounds *in silico* and subsequently validated these findings *in vitro*. This research presents a novel approach to discover lead compounds that target multiple nucleic acid morphologies.

TABLE OF CONTENTS

PAGE

ACKNOWLEDGMENTS....................iii
ABSTRACT....................iv
LIST OF TABLES....................vii
LIST OF FIGURES....................ix

CHAPTER

I. INTRODUCTION AND BACKGROUND....................1

II. MOLECULAR DOCKING OF INTERCALATORS AND GROOVE BINDERS TO NUCLEIC ACIDS USING AUTODOCK AND SURFLEX....................24

III. DEVELOPMENT OF *IN SILICO* PREDICTIVE METRICS THAT GOVERN SMALL MOLECULE-NUCLEIC ACID INTERACTIONS....................81

IV. DISCOVERY OF NOVEL TRIPLE HELICAL DNA INTERCALATORS BY AN INTEGRATED VIRTUAL AND ACTUAL SCREENING PLATFORM....................233

V. DISCOVERY OF A G-QUADRUPLEX NUCLEIC ACID BINDING SMALL MOLECULE BY *IN SILICO* SCREENING AND MOLECULAR MODELING....................261

VI. CONCLUSIONS....................317

REFERENCES....................323
CURRICULUM VITAE....................335

LIST OF TABLES

TABLE PAGE

1. A sampling of the various target classes for which ligands have been successfully identified by computational approaches.............................4
2. Autodock Grid Map Coordinate Dimensions and Grid Center Information..36
3. The following table describes the nucleic acid morphologies, sequences and sites targeted for molecular docking and are identical to those used in competition dialysis...90
4. The classification of compounds from the 67 Compound library set grouped by chemical similarity..98
5. Autodock Grid Properties used for Docking Experiments.....................102
6. The nucleic acids targeted for docking experiments. The targets highlighted in green were used in the original molecular docking experiments while the targets highlighted in yellow were added later to augment the initial experimental design..................105
7. The following table describes the groove binding and intercalation metrics that were developed for Surflex and Autodock...117
8. Comparison of Software for the Prediction of Compound Binding Mechanism After Application of the *In Silico* Metrics...............229

9. Oligonucleotide Array for Competition Dialysis....243

10. Competition dialysis metric results for the positive controls, MHQ-15, OZ-85H, and the virtual screening top results, compounds 1 and 2....253

11. Autodock Docking Parameters used for *In Silico* Targeting of the End-Pasting Sites in Hybrid-1 and Hybrid-2.....284

12. The scores for the top-ranked pose from Autodock or Surflex for the four possible end-pasting sites in Hybrid-1 and Hybrid-2.....312

LIST OF FIGURES

FIGURE PAGE

1. A drawing of the competition dialysis assay setup 13
2. Most Common Anatomical Sites for Cancer Deaths for Males and Females..16
3. Chemical structures of the four test ligands used in the Autodock and Surflex docking studies 29
4. Chemical structure of Furamidine, the ligand used to generate Surflex 2.11 protomols..39
5. Autodock and Surflex accuracy: The dock with the lowest RMSD is presented, regardless of ranking...46
6. The top ranked pose by Autodock and Surflex...............................48
7. The rank of the lowest RMSD pose out of all dockings.........................53
8. Comparison of the top ranked Autodock pose to the PDB crystallographic pose...56
9. Comparison of the top ranked Surflex pose to the PDB crystallographic pose...59
10. Autodock and Surflex Parameterization Accuracy...............................64
11. Autodock and Surflex Parameterization Ranking...............................66
12. Comparison of the Free Energy of Binding for the Crystal pose, Autodock top ranked pose and Surflex top ranked

pose for various ligands using the Autodock and Surflex Scoring Functions..........69

13. Calculated Energy of Binding by Macromodel for the Crystal pose, Autodock top ranked pose and Surflex top ranked pose for various ligands using the Amber and OPLS2005 Force Fields with and without implicit water salvation..........72

14. Some of the nucleic acid structures used for the molecular docking experiments..........92

15. The Positive Control Set of Ligands used for the molecular docking experiments..........95

16. Surflex-Dock (top) and Autodock (bottom) docking scores for the Positive Control set of compounds, before application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library..........107

17. Surflex-Dock (top) and Autodock (bottom) docking scores for the Positive Control set of compounds, before application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library..........109

18. (A) The triplex DNA poly(dA)-[poly(dT)]$_2$ with an intercalation site that is designated as the target site by the Surflex-Dock protomol. (B) The same DNA with a protomol covering all available binding sites including the intercalation site, major groove, minor groove and minor-minor groove..........113

19. Surflex-Dock (top) and Autodock (bottom) docking scores for the Positive Control set of compounds, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library....................120

20. Surflex-Dock (top) and Autodock (bottom) docking scores for the Positive Control set of compounds, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library............................123

21. The Validation Set of Ligands used for the molecular docking experiments...126

22. Surflex-Dock (top) and Autodock (bottom) docking scores for the Validation set of compounds, before application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library..129

23. Surflex-Dock (top) and Autodock (bottom) docking scores for the Validation set of compounds, before application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library...........................131

24. Surflex-Dock (top) and Autodock (bottom) docking scores for the Validation set of compounds, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.........................133

25. Surflex-Dock (top) and Autodock (bottom) docking scores for the Validation set of compounds, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library..........................135

26. The Subsets of the 67 Compound Set of Ligands used for the molecular docking experiments....................................138

27. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the ethidium bromide derivative set from the 67 Compound Set...149

28. Surflex-Dock (top) and Autodock (bottom) docking scores for the ethidium bromide derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library..............................151

29. Surflex-Dock (top) and Autodock (bottom) docking scores for the ethidium bromide derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library...153

30. Surflex-Dock (top) and Autodock (bottom) docking scores for the acridine derivatives from the 67 Compound Set, after application of the groove binding metrics.

The results shown are for the groove binding sites in the nucleic acid library..156

31. Surflex-Dock (top) and Autodock (bottom) docking scores for the acridine derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library..158

32. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the acridine derivative set from the 67 Compound Set.................160

33. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the aromatic diamidine derivative set from the 67 Compound Set...163

34. Surflex-Dock (top) and Autodock (bottom) docking scores for the aromatic diamidine derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library...............................165

35. Surflex-Dock (top) and Autodock (bottom) docking scores for the aromatic diamidine derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the

groove binding sites in the nucleic acid library....................167

36. Surflex-Dock (top) and Autodock (bottom) docking scores for the cyclic aromatic derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library...171

37. Surflex-Dock (top) and Autodock (bottom) docking scores for the cyclic aromatic derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library..173

38. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the cyclic aromatic derivative set from the 67 Compound Set...176

39. Surflex-Dock (top) and Autodock (bottom) docking scores for the dibenzophenanthroline derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library............................179

40. Surflex-Dock (top) and Autodock (bottom) docking scores for the dibenzophenanthroline derivatives from the 67 Compound Set, after application of the

intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library....................181

41. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the dibenzophenanthroline derivative set from the 67 Compound Set..183

42. Surflex-Dock (top) and Autodock (bottom) docking scores for the bis-quinoline derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library..........................186

43. Surflex-Dock (top) and Autodock (bottom) docking scores for the bis-quinoline derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library...................................188

44. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the bis-quinoline derivative set from the 67 Compound Set......190

45. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the amidoanthraquinones (aromatic side chains) derivative set from the 67 Compound Set.................................193

46. Surflex-Dock (top) and Autodock (bottom) docking scores for the amidoanthraquinones (aromatic side chains) derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library......................................195

47. Surflex-Dock (top) and Autodock (bottom) docking scores for the amidoanthraquinones (aromatic side chains) derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library...198

48. Surflex-Dock (top) and Autodock (bottom) docking scores for the amidoanthraquinones (non-aromatic side chains) derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library...200

49. Surflex-Dock (top) and Autodock (bottom) docking scores for the amidoanthraquinones (non-aromatic side chains) derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the

nucleic acid library..202

50. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the amidoanthraquinones (non-aromatic side chains) derivative set from the 67 Compound Set.....................204

51. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the naphthoflavone derivative set from the 67 Compound Set...207

52. Surflex-Dock (top) and Autodock (bottom) docking scores for the naphthoflavone derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library...................................209

53. Surflex-Dock (top) and Autodock (bottom) docking scores for the naphthoflavone derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library..211

54. Surflex-Dock (top) and Autodock (bottom) docking scores for the amidofluorenone derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic

acid library..214

55. Surflex-Dock (top) and Autodock (bottom) docking scores for the amidofluorenone derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library..................................216

56. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the amidofluorenone derivative set from the 67 Compound Set...218

57. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the Other Compound set from the 67 Compound Set...........221

58. Surflex-Dock (top) and Autodock (bottom) docking scores for the Other Compound set from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library...................................224

59. Surflex-Dock (top) and Autodock (bottom) docking scores for the Other Compound Set from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library...................................226

60. Chemical structures of the ligands used in virtual screening and competition dialysis Experiments..............249

61. Competition dialysis results for MHQ-15, OZ-85H compound 1 and compound 2..................................251

62. Induced Circular Dichroism results for (A) compound 1 and (B) compound 2...256

63. Thermal Melting results for (A) compound 1 and (B) compound 2...258

64. (A) A G-quadruplex structure (PDB ID 2JPZ) that contains the human telomeric repeat and shows potential ligand interaction sites in the grooves, intercalation sites and end-pasting sites. (B) The G-quadruplex (PDB ID 1NZM) with the sequence $(TTAGGT)_4$ with a Guanine-Adenine intercalation site that was initially used for virtual screening with Autodock and Surflex....................................268

65. CD Spectrum (A) and CD Melt (B) of the K+ 22mer G-Quadruplex...275

66. The structure of the Quaterpurine small molecule used to "open-up" the Hybrid-1 and Hybrid-2 End Pasting sites...281

67. The virtual screening strategy used to dock the *in silico*

library of compounds to the $(TTAGGT)_4$ quadruplex and determine the best hits for biophysical testing…………..…289

68. The newly discovered small molecule, Compound 1, from the *in silico* virtual screening experiments…………291

69. UV/Vis Absorption Titrations at increasing Quadruplex/Compound 1 ratios demonstrating the G-quadruplex interaction of Compound 1 with the K^+ 22mer quadruplex, $AGGG(TTAGGG)_3$…………294

70. Induced CD Spectroscopy at increasing Quadruplex/Compound 1 ratios showing the G-quadruplex interaction of Compound 1 with the K^+ 22mer quadruplex, $AGGG(TTAGGG)_3$…….298

71. TO-FID results showing the displacement of Thiazole Orange (TO) from the K+ 22mer quadruplex and subsequent TO fluorescence quenching at increasing concentrations of the ligand Compound X……..……………………………...301

72. Effect of Compound 1 on the Melting Temperature of the FAM-TAMRA labeled K^+ 22mer G-Quadruplex……304

73. Effect of a 20 fold Excess of Competitor DNA on the Ability of Compound 1 to Thermally Stabilize the FAM-TAMRA labeled K^+ 22mer G-Quadruplex………..307

74. The top ranked poses for Compound X docked using

Surflex (Gold) and Autodock (Cyan) to the following G-quadruplex nucleic acid structures: (A) Hybrid-1 End Paste Site 1 (B) Hybrid- 1 End Paste Site 2 (C) Hybrid-2 End Paste Site 1 and (D) Hybrid-2 End Paste Site 2...310

75. The top ranked poses for Compound X docked using Surflex (Gold) and Autodock (Cyan) to the following G-quadruplex nucleic acid structures: (A) Hybrid-1 End Paste Site 1 (B) Hybrid-1 End Paste Site 2 (C) Hybrid-2 End Paste Site 1 and (D) Hybrid-2 End Paste Site 2...314

CHAPTER I

INTRODUCTION AND BACKGROUND

Modern day drug discovery has focused almost exclusively on targeting proteins. While these efforts have resulted in many therapeutic successes, other classes of targets such as nucleic acids have largely been ignored. In fact, fewer than 2% of currently marketed drugs and biologicals target nucleic acids [1]. This is most unfortunate as nucleic acids represent promising targets for indications ranging from microbial infections to cancer [2-5]. In the past, this lack of focus on nucleic acids as small molecule targets may be partly ascribed to limited knowledge of the diversity of nucleic acid structure and function. Recently, much scientific progress has been made in the understanding of the physiological relevance of duplex, triplex and G-quadruplex morphologies of nucleic acids and these structures are becoming increasingly attractive small molecule targets [2, 6-8]. Furthermore, various classes of small molecules have been shown to bind to unique nucleic acids in a sequence and structurally specific manner, as has been elegantly demonstrated by Dervan with the hairpin polyamides and Chaires with multiple small molecule families [9-10]. This research has paved the way for the approach of discovering novel small molecules that specifically target newly discovered nucleic acids that may have particular therapeutic or clinical relevance.

Nucleic Acid Structures are Promising Small Molecule Targets

Nucleic acids have long escaped therapeutic targeting because of a lack of knowledge and appreciation of the structural and functional diversity of these macromolecules. It is now known that DNA can have tremendous diversity with respect to structure, conformation and sequence. For example, DNA can exist as a single strand or as duplex, triplex and quadruplex structures. DNA can adopt a large number of secondary and higher order structures *in vivo*, including the standard B-form duplex DNA as well as other duplex structures such as the Z-form duplex DNA. The sequence composition also adds a unique dimension of diversity to DNA. Small molecules have been discovered that may bind to particular DNA structures with moderate selectivity and modulate biological activity *in vivo*. One example is the small molecule telomestatin, which has been shown to bind to G-quadruplex structures with a greater than 70 fold preference compared to duplex DNA and has possible anti-cancer cell activity [11]. This suggests that it is possible to identify small molecules with a preference for specific nucleic acid structures. The discovery of novel small molecules to date appears to be mostly limited to isolation from natural sources and chemical synthesis and sorely overlooks the capability of *in silico* virtual screening and computational approaches.

Virtual Screening Approaches for Discovering New Drugs

In silico virtual screening techniques are valuable computational tools for the discovery of new small molecules that can bind to a target of interest [12]. Indeed, computational methods have been integrated into the discovery process for over 50 compounds that are in clinical trials as well as marketed drugs [13]. Table 1 shows

Table 1. A sampling of the various target classes for which ligands have been successfully identified by computational approaches.

Table 1. A sampling of the various target classes for which ligands have been successfully identified by computational approaches. Adapted from [14].

Target Family	Target Name	Manuscript Reference
Enzyme	Renin	[15]
Drug Metabolizing Enzymes	Cytochrome P450s	[16]
Kinases	Protein Kinase C	[17]
Transporter	Na+/D-glucose co-transporter	[18]
Receptor	AMPA receptor	[19]
Channels	Potassium and Sodium Channels	[20]
Transcription Factors	AP-1 transcription Factor	[21]
Antibacterial	Mycobacterium tuberculosis thymidine monophosphosphate kinase	[22]
Antivirual	Neuroamidase	[23]

compounds that have been discovered using various computational methods against a wide array of target classes emphasizing the importance of *in silico* approaches in discovering new compounds in many research areas.

The benefits of virtual screening are its speed, accuracy, hit rates and affordability, which circumvent the often laborious, slow and expensive process of synthesis of novel small molecules for testing purposes. These benefits have accelerated the adoption of virtual screening in the drug discovery process and it is estimated that up to 20% of new drugs will be found by virtual screening methods in the year 2010 [24]. There are multiple ways to perform *in silico* virtual screening experiments as well as many small molecule databases that can be used for *in silico* screening that will be described in detail below.

Virtual screening experiments are typically considered to be either structure-based or ligand-based [25]. Structure-based virtual screening methods require the availability of an *in silico* structure of the target. This structure is usually obtained through high-resolution X-Ray crystallography techniques or by NMR methods. Some of the most widespread resources for many *in silico* solved structures are the RSCB protein data bank (PDB) and the Nucleic Acid Database (NDB). These databases are popular as the structures can be visualized through a web interface and downloaded directly for virtual screening experiments. Structure-based virtual screening uses various software packages to screen millions of compounds to determine how well each compound can fit into a site on the three dimensional target of interest [26]. This approach involves both "docking" the compounds to the target as well as "scoring" the poses and determining which pose is "correct" [27]. The "scoring" and ranking of the top poses of each ligand in the binding pocket of the target is one of the most challenging aspects of docking [12]. Molecular docking using programs such as DOCK, Autodock, Ludi, FlexX and Surflex-Dock have been used to find many lead molecules against a variety of targets, of which the vast majority are proteins such as thymidylate synthase, retinoic acid receptor, kinases, estrogen receptor and thrombin [14, 28-30]. The use of molecular docking appears well entrenched in academia and industry and its use will likely increase as virtual databases of small molecules and drug targets continue to expand.

A second type of virtual screening approach is referred to as ligand-based virtual screening which requires knowledge of the structure of a biologically active ligand. The structure of the active compound is compared to millions of other chemical compounds to check for chemical and morphological similarity. The premise is that if the structure of the test compound is similar to that of the known active compound, then the test compound may possess similar biological activity [27]. If multiple small molecules are known to possess similar biological activity, a "pharmacophore" can be constructed which describes the ligand chemical properties that are necessary for a ligand to interact with its target. This "pharmacophore" modeling can be particularly useful to detect a wide number of compounds with diverse chemical features [25]. One consideration with ligand-based virtual screening that it does not require knowledge of the structure of the target. This can be advantageous because it can be difficult and sometimes controversial to actually use the "correct" structure of the target for docking studies. However, it is also disadvantageous in that critical interactions of the active compound with the target such as hydrogen bonding and steric interactions may not be effectively visualized and assessed. Ligand-based virtual screening is a popular approach to look for derivatives of known biologically active compounds. This approach has also been used to enrich databases for possible selection of lead compounds [27]. Programs such as FlexS, fFlash and Surflex-Sim have been previously used with success for ligand-similarity based searches [25].

A final aspect of virtual screening is the importance of the repository of small molecules that are used for screening experiments. The database of compounds for virtual screening has increased dramatically in recent years, with tens of millions of compounds currently available in multiple databases [27]. In our own experience, one of the ZINC databases that we use for virtual screening experiments has increased from approximately 2.7 million compounds in 2007 to over 10.6 million compounds in 2009, the vast majority of which are purchaseable from vendors world-wide. The value in having large databases is the large chemical space that these compounds encompass. This vastly increases the number of small molecules considered as possible lead candidates which is favorable compared to the relatively few molecules that are evaluated by actual chemical synthesis and other drug discovery techniques. Additionally many of the *in silico* libraries have been filtered based on specific criteria (for example, Lipinski's Rule of 5) to increase the chance that the molecules are "Drug-Like" in behavior. In the case of Lipinski's Rule of 5, a structural analysis was performed on a large library of drugs that are either currently marketed or in clinical trials. The following rules were developed (coined "Lipinski's Rule of 5") to characterize a small molecule as "Drug-Like" as the vast majority of compounds that were in the library possessed these properties: ≤5 hydrogen bond donors, ≤10 hydrogen bond acceptors, ≤500 daltons molecular weight and ≤5 octanol-water partition coefficient (Log P) [31]. Taken in total, virtual screening against large databases of compounds rationally explores a much larger chemical space than using other approaches such as chemical synthesis and represents a novel way to discover new lead candidate small molecules against a target of interest.

Virtual screening targeting DNA forms has been largely ignored

While the use of virtual screening for the discovery of new ligands that target proteins has been well established, very few studies have been performed with nucleic acids [2, 32]. This may be partly because almost all virtual screening software has been designed for proteins, and may not account for characteristics that are particularly important to nucleic acids such as their distinct geometrical symmetry and the electrostatic effects of the phosphate backbone. Moreover, there are few published reports of the use of these programs to target nucleic acids [33-34]. Perhaps the greatest gap in knowledge in this area is the lack of a systematic study to determine whether docking software can accurately reproduce known crystal structures of ligands bound to nucleic acids and also predict the binding mechanisms of small molecules to nucleic acids, which we address here.

Small molecules typically interact with duplex nucleic acids by binding to the minor groove or by intercalation between existing base pairs [4, 10, 35]. The geometry of the grooves of triplex and quadruplex structures may have structural features that make these nucleic acids unique compared to the major and minor grooves of duplex B-DNA. The quadruplex structures in particular have diverse loop regions that may be functional targets for small molecule binding. It is of primary interest to develop virtual screening metrics that can differentiate small molecules that bind by either minor groove binding or intercalation. This is important because correctly predicting the nucleic acid structural selectivity and binding mechanism of small molecules is critical for understanding the therapeutic potential and non-specificity of a ligand. It remains of paramount importance to first, ascertain whether molecular docking software can be used to target nucleic acids and second, if novel rules can be developed to predict nucleic acid structural selectivity and the binding mechanism of a given small molecule. This will serve dual roles in filling a major basic science knowledge gap in predicting how small molecules bind to nucleic acids and also provide potentially enormous opportunities for translating this knowledge into the discovery of new therapeutic small molecules.

Limitations in Previous Virtual Screening Studies

A limited number of virtual screening studies against nucleic acids suggest that it is possible to successfully target these structures for small molecule discovery. The DOCK program in particular was used by Grootenhuis and Chen to target duplex DNA and RNA, respectively [33, 36-37]. Rohs *et al.* used a Monte Carlo algorithm to assess binding of methylene blue to DNA [38]. Shafer and Kuntz discovered a carbocyanine dye (DODC) that binds to G-quadruplexes [39]. Finally, Evans *et al.* appears to have one of the most comprehensive studies assessing minor groove binders to DNA using Autodock [34]. However, the Evans study was limited and did not assess ligands that bind by intercalation and did not exhaustively explore the Autodock parameters, which can significantly affect docking performance and outcome. While all of these studies suggest it is possible to use virtual screening to target nucleic acids, none of the studies comprehensively compared the ability of the software to reproduce multiple minor groove binder and intercalator crystal structures or assessed the software for large scale virtual screening feasibility. A major deficiency of the studies is a lack of a knowledge base for *in silico* prediction of the mechanism of action of a ligand.

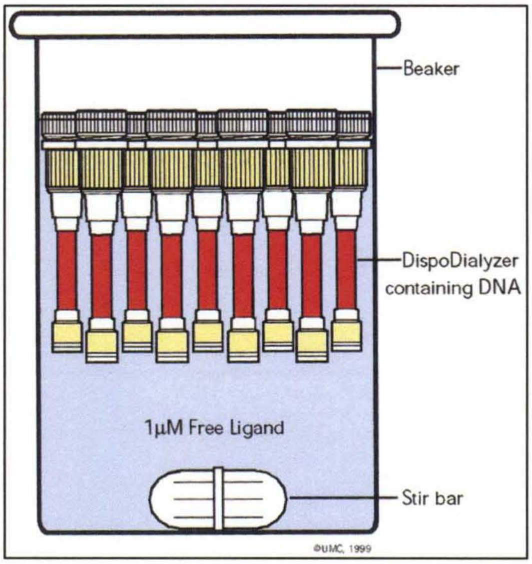
Experimental Validation of Predicted *In silico* "Hits"

A necessary complementary technique to any virtual screening approach is empirical testing of the "hits" that are identified from the *in silico* virtual screen. This is important to distinguish the false from true positive hits from the *in silico* screening data [27]. There is much debate about which techniques are appropriate for assessing the interaction of a small molecule with an array of nucleic acids. Several methods include ESI-MS (Electrospray Ionization Mass Spectroscopy), FRET-melting (Flourescence Resonance Energy Transfer), SPR (Surface Plasmon Resonance), Fluorescence Intercalation Displacement Assay (FID) and competition dialysis [40]. The method of competition dialysis is preferred as it has distinct advantages over the others, although the methods of FID has advantages as well and is complementary to competition dialysis. For example, ESI-MS requires changing the salt condition of the nucleic acid out of sodium and potassium and typically into ammonium acetate, which may dramatically impact the structure of nucleic acid morphologies, particularly the therapeutically relevant quadruplex structures [40-43]. FRET-melting suffers from having to modify the oligonucleotides with a fluorescent probe and possible ligand-probe fluorescence interference [40]. Finally, while SPR has the advantage of high sensitivity in assessing small molecule-nucleic acid interactions, either the ligand or nucleic acid must typically be covalently modified and bound to a chip for analyzing the interaction, as opposed to allowing the interaction to occur free in solution [40]. Additionally, great expertise is required in choosing the appropriate chip for assessing the interactions as well as significant capital expenditure in purchasing the instrument. For these reasons, ESI-MS, FRET-melting and SPR techniques have substantial limitations for assessing the ligand-nucleic interactions as described here. FID is complementary to competition dialysis and may have particular utility if a small molecule lacks a suitable chromophore for competition dialysis testing. The assay relies upon the known intercalation of a reporter dye such as ethidium bromide or thiazole orange into a DNA of interest. The fluorescence of such reporter molecules is markedly increased upon binding to the nucleic acid and quenched when free in solution. Thus, the assay can be used for competition experiments where small molecule can be added to a solution containing DNA and thiazole orange and the fluorescence of thiazole orange can be monitored to determine if it is bound or displaced from the DNA. We describe in more detail the use of this assay for characterizing the binding mode of some newly discovered compounds in Chapter V.

On the other hand, competition dialysis is a simple, rapid technique that has gained world-wide acceptance as a way to quantitatively and rigorously assess the binding of small molecules to nucleic acids [10]. The assay can determine the sequence and structural selectivity of a single ligand for any nucleic acid of interest. The setup involves dialyzing a set of nucleic acids at identical concentration against a common dialysate containing the ligand of interest. As the system reaches equilibrium, the ligand will accumulate in the dialysis cassette containing the nucleic acid to which the ligand

Figure 1. A drawing of the competition dialysis assay setup.

Figure 1. A drawing of the competition dialysis assay setup. Adapted from [44].



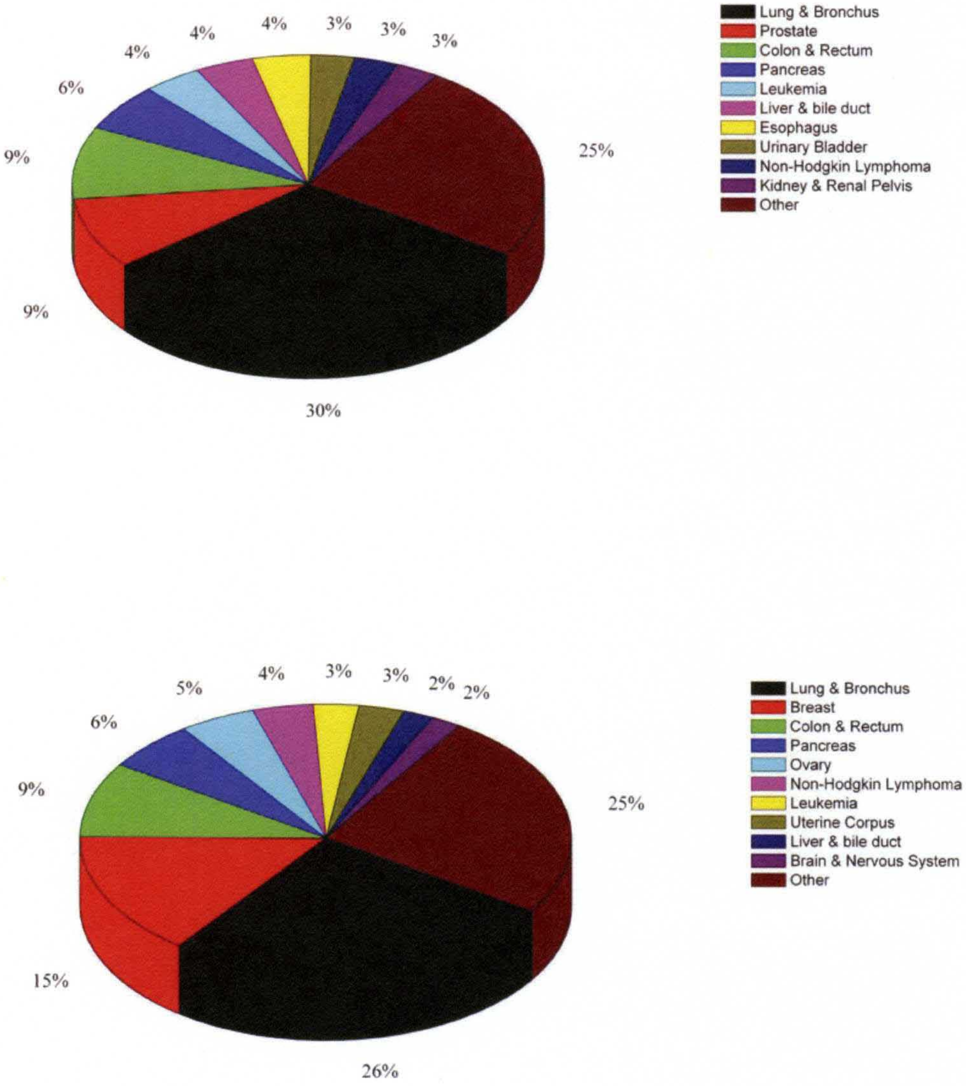
binds the tightest [10]. The ligand is then dissociated from the nucleic acid using detergents and quantified by either absorbance or fluorescence. The current version of the assay typically consists of 19 nucleic acid species including duplex, triplex and G-quadruplex morphologies. However, the original 19 structures is but a starting point for the assay. The power of this assay is the customizability and freedom of choice of the nucleic acid structures; essentially any unique nucleic acid sequence or morphology can be added to the array of nucleic acids and tested for ligand binding. Additionally, the technique allows for a comparison of the ligand binding properties for many nucleic acids that are simultaneously free in solution. This highlights the substantial benefit of this technique compared to the previously mentioned methods. Competition dialysis has proven valuable in assessing ligand affinity and selectivity for any nucleic acid species and has particular utility as described here for testing the binding behavior of a small molecule that is predicted from virtual screening metrics.

In silico Discovery of Novel Small Molecules with Therapeutic Potential

The ultimate goal in our research is to combine our *in silico* research with actual testing by competition dialysis and other techniques to provide an integrated platform to discover new small molecules that bind to physiologically important nucleic acids. The determination of predictive metrics for the purposes of discovering novel small molecules that can bind nucleic acids could have substantial therapeutic benefit in many areas of disease, most notably cancer. The CDC estimates from 2006 placed cancer as the second leading cause of death in the United States, second only to cardiovascular disease. Most recently in 2008 in the United States alone, an estimated 565,650 people succumbed to cancer [45] which can affect many different organ systems (Figure 2). In fact, the

Figure 2. Most Common Anatomical Sites for Cancer Deaths for Males (Top Figure) and Females (Bottom Figure). Adapted from [46].

Figure 2. Most Common Anatomical Sites for Cancer Deaths for Males (Top Figure) and Females (Bottom Figure). Adapted from [46].



lifetime probability of a male developing cancer is 1 in 2 and 1 in 3 for females [47]. As Figure 2 shows, cancer can arise in many anatomical positions and is a major cause of morbidity and mortality in the United States. In recent years, the scientific and medical community has developed new cancer drugs in response to the demand for new treatments. A substantial number of cancer drugs have been approved that are now considered essential for treating various forms of cancer. In particular, biologicals such as monoclonal antibodies have become attractive treatments for specific cancers because of their remarkable specificity and minimal adverse effects [48]. An example is Cetuximab, an Epidermal Growth Factor Receptor (EGFR) Inhibitor, which is approved for the treatment of locoregionally advanced squamous-cell carcinoma of the head and neck (LASCCHN) [49]. While biologicals such as Cetuximab have undoubtedly benefited patients, these large molecules are costly and time-consuming to manufacture and the cost is prohibitive for many patients.

Even though the vast majority of new cancer treatments are focused on protein targets, there are some existing therapeutics that work by targeting nucleic acids. The anthracylines, for example, have been a key class of drugs that target DNA for cancer chemotherapy for over 40 years, despite suffering from severe side effects [50-51]. An example is cisplatin which is a chemotherapy drug that induces cross linking of DNA and is indicated for the treatment of various sarcomas and head and neck cancers. Unfortunately, the major limitation of current nucleic-acid based therapies such as cisplatin is target non-specificity and toxic side effects, which include in the case of cisplatin, severe ototoxicity and neurotoxicity [52]. The development of new anti-cancer drugs based on nucleic acid targets has stagnated until recently.

A new area of cancer drug development is in the area of G-quadruplex nucleic acid structures. These quadruplexes have been observed in the human telomeric region of chromosomes and have a novel mechanism of possibly inhibiting cancer cells replication [53]. Since over 85% of cancer cells overexpress the reverse transcriptase enzyme telomerase, cancer cells are able to maintain the human telomere sequence $(TTAGGG)_n$ which is responsible for cancer cell immortality [54]. G-quadruplex structures have been shown to destabilize telomerase from the telomere, resulting in decreasing cancer cell life [55]. Thus, these quadruplexes have become a source of great interest for the identification of highly selective, small molecules that may bind and stabilize the structures *in vivo*, and inhibit telomerase activity. In fact, there are several G-quadruplex interacting small molecules currently in clinical trials including Quarfloxin (Cylene Pharmaceuticals). This area is one of the most promising areas of current anti-neoplastic small molecule development. As we will describe next in the Dissertation Overview, we target tetraplex nucleic acids to test the therapeutic utility of these novel, predictive, virtual screening metrics. Additionally, the morphologically distinct triplex nucleic acids are targeted because of their ability to potentially modulate gene expression [56-58]. Targeting of triplex and tetraplex nucleic acid structures will demonstrate the power and utility of this new scientific knowledge for the identification of small molecules that can selectively bind to these targets.

Dissertation Overview

In silico virtual screening approaches have been under-utilized for small molecule discovery because of the inability to predict how small molecules interact with nucleic acids. There is a clear need to determine if this behavior can be predicted *in silico* and validated *in vitro*. To meet this need, the goal of this research is to determine if rules can be developed to predict the binding behavior of novel small molecules to therapeutically relevant nucleic acids.

The first goal of this research as detailed in Chapter II is determining if virtual screening methods can be used for targeting nucleic acid structures. Two software packages Surflex and Autodock, are selected for the purposes of validating nucleic acids as feasible targets. Autodock is selected because it is one of the most widely cited molecular docking software [59]. Surflex is chosen since it has the proven advantage of rapid docking which may have particular utility for large scale virtual screening applications [60]. This is a key initial step in this research, as it must be determined if the currently available software is appropriate for evaluating small molecule interactions with nucleic acids. Four nucleic acid-ligand structures were chosen that represent the two major mechanisms (minor groove binding and intercalation) that small molecules use to bind to nucleic acids. The anti-malarial drug pentamidine and the antiviral drug distamycin are two well known drugs that bind to the minor groove of duplex nucleic acids [5, 61]. Daunorubicin and ellipticine are anti-neoplastic drugs that were selected as prototypical nucleic acid intercalators [62]. We demonstrate that both Autodock and Surflex are able to accurately reproduce the *in silico* structures of these ligand-nucleic acid complexes. Interestingly, the docking results change dramatically with the various paramaters that can be customized with the software. The "optimal" parameters for balancing docking accuracy and ranking were determined and serve as the basis for the software operation for the remaining chapters of this dissertation. The results of the work support the use of Surflex in particular for virtual screening applications as the software was found to be approximately 10 fold faster than Autodock with comparable docking accuracy and ranking. Some considerations and limitations of the software are also detailed. The results of this work are published in P.A. Holt *et al* [63].

After demonstrating that molecular docking software can reproduce multiple known ligand-nucleic acid crystal structures, the focus of Chapter III is on whether rules can be developed to predict the nucleic acid structural specificity and binding mechanism of a ligand. This is significant as a major hurdle to current drug development is small molecule non-specificity, which can result in drug toxicity and significant adverse effects. An *in silico* nucleic acid library with 10 structures was constructed including duplex, triplex and quadruplex morphologies of nucleic acids with appropriate groove binding and intercalation sites for docking the ligand. The small molecules from Chapter II (daunorubicin, distamycin, ellipticine and pentamidine) were docked to the compounds and *in silico* rules were developed to classify the binding mechanism and sequence selectivity of these molecules, based on their known binding behavior. The rules were tested on several triplex and quadruplex binding ligands that our lab has recently discovered as well as on a set of 67 minor groove binder and intercalator compounds that have been previously tested by competition dialysis [10]. The results showed that the metrics were able to generally accurately predict whether the compounds were groove binders or intercalators, but predicting sequence specificity was more challenging. In general, Surflex appeared to outperform Autodock and appears more appropriate for large scale virtual screening efforts.

The knowledge gained from Chapters II and III is utilized for the discovery of new ligands that can bind to a specific nucleic acid and this work is described in Chapter IV. In this chapter, both ligand and structure based virtual screening techniques are combined as well as utilizing the established *in silico* selectivity metrics to discover new triplex nucleic acid binding small molecules. Chaires *et al.* have previously identified a set of napthylquinoline ligands that were demonstrated by competition dialysis to be highly selective triplex poly(dA)-[poly(dT)]$_2$ intercalators [44]. One of these napthlyquinolines in particular, MHQ-12, was used as the parental ligand in a similarity search against millions of *in silico* compounds. For the top similarity hits, additional structure-based docking studies were performed and *in silico* selectivity metrics were applied. Two novel compounds were discovered that were tested by competition dialysis, UV/Vis thermal melting and circular dichroism and were demonstrated to be highly selective intercalators into the targeted triplex DNA. This demonstrated the practical application of the *in silico* metrics that were discovered in the previous chapter and shows that novel small molecules can be discovered using an integrated *in silico* and biophysical testing platform. The results of this work are published in P.A. Holt *et al* [64].

Chapter V focuses on the structure-based targeting of G-quadruplex nucleic acids for the purposes of discovering new small molecules. The work details the targeting of the AGGG(TTAGGG)$_3$ G-quadruplex which is found with increasing frequency in the single stranded overhang of the human telomeric region of chromosomes. Using the previously optimized software parameters, Surflex and Autodock were used to screen over 6.6 million compounds that may interact with the G-quadruplex. Ligands that bind by intercalation or at the end of quadruplexes (by "end pasting") are particularly appealing as they may stabilize the quadruplex structure by interactions with the guanine quartets. Stabilization of G-quadruplexes can dissociate telomerase and result in decreased cancer cell proliferation [55]. A consensus scoring approach was applied which combines the top scoring results for Surflex and Autodock and re-ranks the results. The top compounds were tested by spectroscopic and fluorescent methods and a compound was discovered that interacts with the G-quadruplex DNA by the hypothesized binding mechanism. Moreover, the scaffold is unlike any reported to date in the literature. The work in this chapter is a practical application of the knowledge discovered in previous chapters and demonstrates that the software and approach as developed in this work, is capable of discovering new small molecules that bind to a nucleic acid by a specific mechanism.

Summary

While nucleic acids represent a viable class of drug targets for *in silico* virtual screening, progress has been hampered by the lack of virtual screening rules that can predict the binding mechanism of a ligand to a nucleic acid target. The development of predictive rules is an essential step to discover novel small molecules to fight disease. It is also a critical part in an integrated virtual and actual screening platform that can screen millions of compounds *in silico* and biophysically test the most promising compounds identified from the initial computational screen. While there has been much progress in the research and understanding of nucleic acids, the therapeutic development of targeting nucleic acids lags behind. This appears to be due to a lack of a rapid, efficient and economical approach to identify selective small molecules that can bind to nucleic acids.

Determination of predictive rules, as described herein, addresses this knowledge gap by making it possible to better understand and predict the interaction of small molecules with nucleic acids. We believe this new information will ultimately facilitate the discovery of novel ligands that target therapeutically relevant nucleic acids.

CHAPTER II

MOLECULAR DOCKING OF INTERCALATORS AND GROOVE BINDERS TO NUCLEIC ACIDS USING AUTODOCK AND SURFLEX

This chapter describes the validation of selected virtual screening software for the purposes of targeting nucleic acids. We demonstrate here that the molecular docking tools Autodock and Surflex accurately reproduce the crystallographic structures of a collection of small molecule ligands that have been shown to bind nucleic acids. Docking studies were performed with the intercalators Daunorubicin and Ellipticine and the minor groove binders Distamycin and Pentamidine. Autodock and Surflex dock Daunorubicin and Distamycin to their nucleic acid targets within a resolution of approximately 2 Å, which is similar to the limit of the crystal structure resolution. However, for the top ranked poses, Autodock and Surflex both dock Ellipticine into the correct site but in a different orientation compared to the crystal structure. This appears to be partly related to the symmetry of the target nucleic acid, as Ellipticine is able to dock from either side of intercalation site but also due to the shape of the ligand and docking accuracy. Surflex docks Pentamidine in a symmetrically equivalent orientation relative to the crystal structure, while Autodock was able to dock this molecule in the original orientation. In the case of the Surflex docking of Pentamidine, the initial RMSD is misleading, given the symmetrical structure of Pentamidine. Importantly, the ranking functions of both of the programs are able to return a top pose within approximately 2 Å RMSD for Daunorubicin, Distamycin and Pentamidine and approximately 3 Å RMSD for Ellipticine compared to their respective crystal structures.

Finally, we also discuss some docking challenges and potential pitfalls when using these software tools, such as the importance of hydrogen treatment on ligands as well as the scoring functions of Autodock and Surflex. Overall for this set of complexes, Surflex is preferred over Autodock for virtual screening, as although the results are comparable, Surflex has significantly faster performance and ease of use under the optimal software conditions tested. These experiments show that the molecular docking techniques can be successfully extended to include nucleic acid targets, a finding which has important implications for virtual screening applications and in the design of new small molecules to target therapeutically relevant morphologies of nucleic acids. The results and conclusions of this scientific research were published by P.A. Holt *et al* [63].

MOLECULAR DOCKING OF INTERCALATORS AND GROOVE BINDERS TO NUCLEIC ACIDS USING AUTODOCK AND SURFLEX

Patrick A. Holt, Jonathan B. Chaires, John O. Trent. Molecular Docking of Intercalators and Groove-Binders to Nucleic Acids using Autodock and Surflex. J Chem Inf Model. 2008 Aug; **48** (8):1602-15. Epub 2008 Jul 22.

Introduction

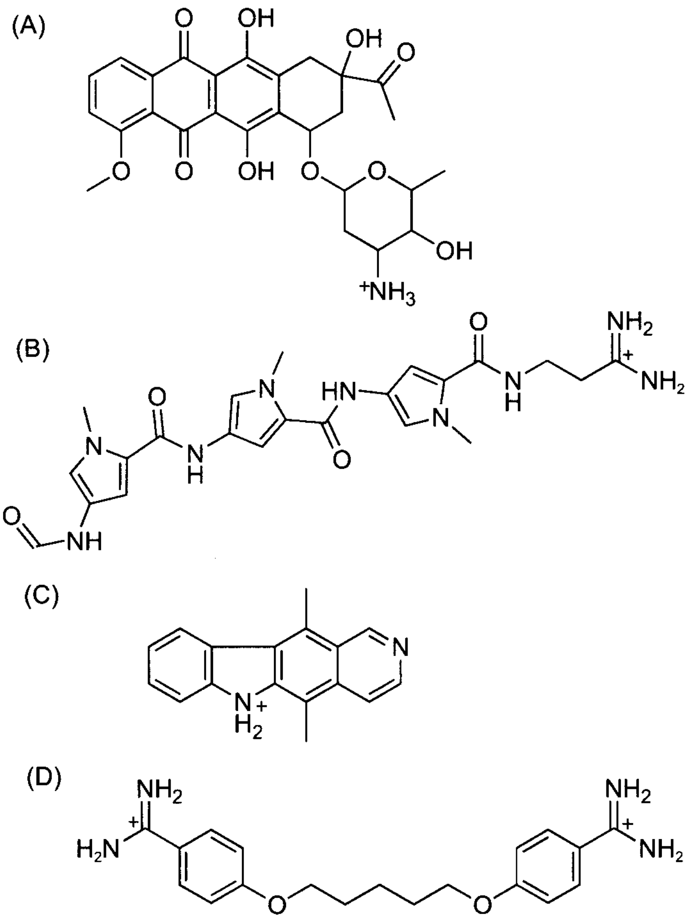
Molecular docking techniques have shown great promise as a new tool in the discovery of novel small molecule drugs for targeting proteins [60, 65-67]. Fewer molecular docking studies have been performed targeting nucleic acids structures, despite advances in the understanding of the functional importance and the unique structural features of duplex, triplex and G-quadruplex morphologies [2, 6-8, 32]. This is unfortunate since not only are there clinically used drugs that target nucleic acids, but many forms of nucleic acids are becoming an increasingly attractive target for anti-neoplastic and anti-microbial agents [2-5, 10, 44, 61, 68-70]. The few docking studies in which nucleic acids are targeted have focused on such sites as the minor groove of DNA, a tetraloop structure of RNA and the major groove of an RNA duplex, while rarely targeting intercalation sites which also hold therapeutic potential [33-34, 36-37, 71-72]. The use of molecular docking has important implications for the synthesis and development of small molecule drugs that selectively target nucleic acids since these techniques have the potential to shed light on the interaction and mechanism of action of these ligands with targets that may have medicinal value.

Small molecules can interact with nucleic acids at multiple sites to alter nucleic acid function [71, 73-74]. In the case of duplex DNA, one drug class binds within the minor groove and a second class intercalates between existing base pairs of the nucleic acid structure [4, 10, 35]. Intercalators and groove binders have distinct thermodynamic signatures that indicate different driving forces for binding [75]. The minor groove is an attractive target for small molecules since this site has less competition from proteins and polymerases, which typically interact with the major groove [5]. An exeption are histone tails which can bind in the minor groove of DNA. The closer proximity of the strands in the minor groove compared to the major groove allows more contact surface area for a small molecule to bind tightly [76]. The unfavorable geometry of the major groove is another reason why few drugs target this groove [71]. Two well-known minor groove binders are the anti-malarial drug Pentamidine and the antiviral drug Distamycin, which we selected for our studies [5, 61, 77-78]. While only limited docking studies have been performed with minor groove binders, even fewer studies have tested whether drugs that act through intercalation can be modeled successfully using docking methods [4, 66]. We selected two prototypical intercalators, Daunorubicin, a drug commonly used to treat certain forms of leukemia, and Ellipticine, another anti-neoplastic drug, for docking experiments using Autodock (4) [79] and Surflex (2.11) (Figure 3) [80].

Autodock 4 and Surflex 2.11 have been used previously for protein-ligand docking, but very few studies have been performed using nucleic acids as targets [2]. Autodock is a logical selection for further exploration as it has been shown in some cases to be superior to DOCK, FlexX and GOLD at reproducing the crystallographic pose of ligand-protein complexes [81]. Surflex was chosen because it has rapid computational

Figure 3. Chemical structures of the four test ligands used in the Autodock and Surflex docking studies. (A) Daunorubicin, (B) Distamycin, (C) Ellipticine and (D) Pentamidine.

Figure 3. Chemical structures of the four test ligands used in the Autodock and Surflex docking studies.



speed with protein-ligand docking which could prove useful for virtual screening [60]. Autodock and Surflex have important differences in search algorithms and scoring functions. A search algorithm is initially used for conformationally sampling the ligand and target interactions, and scoring functions are used for evaluating and ranking the final poses of the ligand to determine the "correct" pose [82].

Autodock performs molecular dockings by pre-calculating energy grids around a site of interest on the target [83]. A stochastic search algorithm utilizing the Lamarkian Genetic Algorithm (LGA) for exploring the grid space is used to perform energy evaluations of the position of the ligand with respect to the target energy grids [83]. This algorithm explores the various orientations and conformations of the whole ligand relative to the energy grids for the defined number of energy evaluations and returns the lowest energy conformation in the target site [83]. The LGA has found particular utility in modeling systems with large numbers of rotatable bonds and possible numbers of conformations [83]. Surflex uses a so-called "whole" molecule alignment algorithm based on morphological similarity between the ligand and target [60]. This docking approach aligns the ligand to a "protomol" or idealized ligand in the active site of the target [60]. The protomol is composed of a collection of fragments or probe molecules that characterize the surface morphology of the binding site [84]. These probe molecules consist of CH_4, C=O and N-H fragments that model steric effects in the binding pocket, hydrogen bond acceptor groups and hydrogen bond donor groups, respectively [60, 84]. The docking ligand fragments are checked for alignment and similarity against the protomol probes [60]. This is referred to as a "whole" molecule approach because after the initial ligand fragmentation, both the small fragment and the rest of the "whole" ligand are carried into the protomol binding site [60]. However, only the small fragment is checked for similarity and alignment against the protomol, while the rest of the "whole" ligand is assessed for steric interactions in the target site after optimal alignment of the fragment [60]. This "whole" molecule approach is powerful because it considers the subsequent position of the rest of the "whole" molecule with respect to the target after the small fragment is optimally aligned with the protomol [60]. This is an important difference between Autodock and Surflex, since Autodock involves evaluation of the conformations of the whole ligand without ligand fragmentation [83].

The scoring functions for Autodock and Surflex are partially empirically based, with Autodock incorporating an Amber type force field and Surflex calculating atom to atom pairwise interactions between the ligand and target [60, 83, 85]. Autodock evaluates pairwise interactions based on van der Waals radii of the atoms to determine the free energy of binding and returns the optimal lowest energy docked conformation as the best docked pose [82]. The Surflex scoring function is parameterized by calculating van der Waals distances between protein and ligand and parameterization of the scoring function was based on 34 protein-ligand complexes [86]. Surflex assigns the atoms as either polar or non-polar and then calculates a score based on hydrophobic and polar contacts between the two atoms [67]. The docked poses are then ranked according to the maximal Surflex Overall score.

Aside from the algorithmic differences in Surflex and Autodock, there are several other aspects of molecular docking in general and these programs specifically that present challenges to successful docking of ligands to nucleic acids. First, because proteins have attracted the most interest as drug targets, proteins have also been the focus of most docking efforts compared to nucleic acids [82]. This leads to the question of whether these protein-configured docking programs will work for nucleic acids because of the unique structural features of nucleic acids including their high charge density, exposed binding sites, and distinct geometrical symmetry [82, 87]. Another challenge is the dependence on crystal structures for visualizing how ligands interact with their targets and for assessing the accuracy of docking software. This approach relies on both the availability and resolution of the crystal structure. For nucleic acids, there are few crystal structures of ligand-nucleic acid complexes available and even small variations in the resolution of the atomic positions of the crystals can significantly affect the modeling of important forces between the ligand and target such as hydrogen bonding [88]. Differences in scoring functions also present a challenge for docking, as ranking of the poses is typically the most difficult aspect of docking [24, 89]. The coefficients and weighting for the scoring function terms are calibrated based on ligand-protein complexes, and it is unknown how well Autodock and Surflex would perform with ligand-nucleic acid complexes [85]. Autodock and Surflex include entropic contributions by accounting for conformational and tortional changes as well as a term for solvation [60, 85]. However, the entropic contribution of solvation terms for most docking programs has been difficult to incorporate accurately in scoring functions and could contribute to erroneous pose ranking [82]. Another traditionally challenging area for docking programs is accounting for target flexibility, since even small conformational changes of the ligand in the binding pocket can cause dramatic changes in the scoring function [67]. While Autodock has the option to explore side chain flexibility for protein receptors, this function has not been extensively explored in the published literature for nucleic acids. Moreover, Surflex does not take target flexibility into account during molecular docking [60]. To fairly compare the performance of these two programs, target flexibility was not considered in these experiments. These are important considerations when performing docking of ligands to nucleic acids using Autodock and Surflex, and could significantly impact docking performance.

In spite of these challenges, however, we demonstrate that Autodock and Surflex can accurately dock small molecules with different binding modes to nucleic acid targets. More importantly, the ranking of the poses is also evaluated, which has been the more challenging aspect for many docking programs [24, 89]. The minor groove binders Distamycin and Pentamidine, and the intercalators, Daunorubicin and Ellipticine were selected for docking studies since these small molecules have crystal structures that are available in the Protein Data Bank (PDB). Autodock and Surflex software operating parameters were evaluated to determine which parameters increase docking accuracy and the successful ranking of the poses. Given the challenge of docking to nucleic acids, some reasons for suboptimal docking are detailed, including the importance of hydrogens on ligands, the scoring functions of the programs, and the quality of the crystal structure. This collection of experiments demonstrate the utility of these programs for molecular docking of ligands to target nucleic acids.

Experimental and Computational Methods

Virtual Library Preparation. Ligand-nucleic acid complex crystal structures for Daunorubicin, Distamycin, Ellipticine and Pentamidine were obtained from the Protein Data Bank with identification numbers of 152d, 2dnd, 1z3f and 1d64, respectively. The resolutions of these structures are 1.4 Å, 2.2 Å, 1.5 Å and 2.1 Å, respectively. Distamycin and Pentamidine are bound to the minor groove of DNA duplex dodecamers d(CGCAAATTTGCG)$_2$ and d(CGCGAATTCGCG)$_2$, respectively. Daunorubicin and Ellipticine intercalate between the Cytosine and Guanine nucleotides in the sequence d(CGATCG)$_2$. For the Ellipticine intercalation PDB structure, Maestro (8.0) [90] was used to construct the symmetrical strand to form a complete, complementary, double stranded DNA. For the intercalator nucleic acid targets, there were two intercalation sites on the target. Thus, the 3' terminal Guanine residue was removed from the 6 base pair sequence so that there would only be a single intercalation site in the target nucleic acid structure. The ligand and nucleic acid targets were saved as separate files for docking purposes.

The PDB files were visually inspected using Macromodel (7.0) [91] and all water molecules were removed. Amber ligand atom types were assigned using Sybyl (7.3) [92] and hydrogen atoms were added as appropriate. The program Antechamber in the software suite Amber (8) [93] was used to assign AM1-BCC charges to the atoms in each of the ligands and to also convert the files from PDB format to MOL2. Python scripts were used to prepare the nucleic acid structures in PDBQT format with Gasteiger charges for use in Autodock experiments while MOL2 files were used for Surflex experiments.

Autodock 4 Methods. Autodock 4 and the graphical user interface Autodock Tools (1.4.6) [94] were compiled for a Macintosh OS X PowerMac G5 and Linux workstations. Autodock Tools 1.4.6 was used for establishing the Autogrid points as well as visualization of docked ligand-nucleic acid structures. The target site on the nucleic acid was specified to encompass either the entire minor groove or the intercalation target site.

Table 2. Autodock Grid Map Coordinate Dimensions and Grid Center Information

Table 2. Autodock Grid Map Coordinate Dimensions and Grid Center Information

Test Ligand	**Grid Point Characteristics (Dimensions)**				**Grid Center Characteristics (Dimensions)**		
	X	**Y**	**Z**	**Total Number of Points**	**X**	**Y**	**Z**
Daunorubicin	52	42	28	66091	14.332	13.212	5.489
Distamycin	34	50	64	98175	9.776	21.55	76.162
Ellipticine	58	32	40	79827	0.992	19.28	46.762
Pentamidine	34	54	52	102025	10.298	20.854	8.457

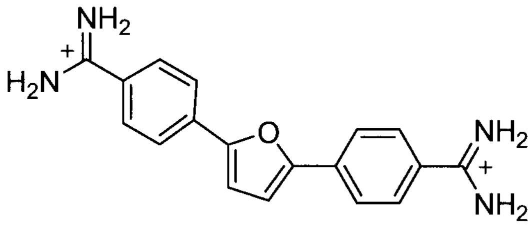
The grid center was also established by centering the grid box on either the minor groove or the intercalation site. The grid maps had a spacing of 0.375 Å (Table 2).

Several available docking parameter options in Autodock 4 were systematically varied to determine the optimal conditions for ligand-nucleic acid docking. These factors include the number of total energy evaluations per docking run and also the total number of docking runs performed. The total number of energy evaluations is the total number of ligand-target energy interaction evaluations before the lowest energy conformation is selected. These factors are suggested as logical starting areas of optimization as they have previously been shown to impact ligand-protein docking studies [95]. The number of energy evaluations per docking run was varied as 200,000 (2E5), 2,000,000 (2E6) or 20,000,000 (2E7). Docking runs were varied as 5, 10 or 20 runs. Thus, a total of nine experiments were performed with varying numbers of energy evaluations and dockings to determine if these factors would impact docking accuracy and ranking. All other docking parameters were left at the default values. For the Autodock parameterization testing experiment with 50 docks and 5E7 energy evaluations, the "ga_num_generations" was set at 100,000. Normally, the docking run will terminate when either the "ga_num_generations" or the number of energy evaluations is reached, so the ga_num_generations was increased to from 27,000 to 100,000 to ensure that 5E7 energy evaluations was reached for these docking experiments [59].

Surflex 2.11 Methods. Surflex 2.11 was compiled for a Macintosh OS X PowerMac G5 and Linux workstations. The protomol was generated using a ligand-based approach, where a small molecule is selected that fits into the site of interest. The structure of the molecule in the site is then used for protomol generation. The protomol represents a set of

Figure 4. Chemical structure of Furamidine, the ligand used to generate Surflex 2.11 protomols.

Figure 4. Chemical structure of Furamidine, the ligand used to generate Surflex 2.11 protomols.



molecular fragments that characterizes the active site and to which the ligand of interest is fragmented and checked for both similarity and alignment [67]. Furamidine was chosen as the ligand for protomol generation, as it has been previously shown to be a minor groove binder and is small enough to fit into the intercalation site to ensure adequate protomol generation (Figure 4) [4, 61, 96]. Importantly, this also reduces the bias of the evaluation by not using the actual ligands to be docked and is a more realistic, generalized docking approach. Two important factors that can significantly effect the size and extent of the protomol generated are "proto_thresh" and "proto_bloat" options. "Proto_thresh" determines how far the protomol extends into the concavity of the target site while "proto_bloat" impacts how far the protomol extends outside of the concavity [97]. For the purposes of these experiments, "proto_thresh" was set to 0.2 and "proto_bloat" was left at the default (0) for all protomols generated except for Daunorubicin, where a "proto_bloat" of 0.5 was used. Protomols were visualized with Sybyl 7.3 to ensure proper coverage of the desired target area.

Surflex 2.11 offers many parameters that can be customized to help optimize ligand targeted docking. An investigation of all of the combinations of these factors is beyond the scope of this paper. Instead, two factors, the "Multistart 5" and "Random 5" options, were selected as these are thought to have the potential to most significantly impact the accuracy of the docked poses. The "Multistart 5" designation enables docking to begin from 5 different initial starting positions around the designated target. Previously, Jain *et al.* had observed little increase in successful docks with protein targets beginning at a value of 5 ("Multistart 5"), relative to the additional computational resources required for docking these extra conformations [97]. A "Random 5" option ensures that the ligand adopts 5 random X,Y,Z coordinate conformations prior to initiating docking calculations. These options are both thought to be important since it minimizes the chance that the ligand may be randomly assigned to an energetically or conformationally unfavorable position from which it cannot recover during the docking. A total of three experiments were subsequently performed, with the first having default Surflex 2.11 options ("No Multistart", "No Random"), the second with implementation of "Multistart 5" and the last experiment with implementation of both "Multistart 5" and "Random 5" to test for a potential synergistic effect between these two options. All other parameters were left at the default values.

RMSD Calculations. One metric for evaluation of the quality of docking results is the difference in the X,Y,Z coordinates between the docked pose and the known crystal structure which can be used to calculate the Root Mean Square Deviation (RMSD) between the two poses. For consistency in evaluation of docked poses, the Surflex 2.11 software RMSD method was used for calculation of the RMSD differences for both Autodock and Surflex results based on only the heavy atoms. This method determines the RMSD between the docked pose and the crystallographic structure using a direct atom to atom comparison of the two structures. An additional Surflex RMSD function (Actual RMSD ISO) was used to account for internal ligand symmetry. This function is independent of atom numbering and computes isomorphisms between the crystal and docked poses, returning the lowest symmetrical RMSD value [98]. The practice of accounting for ligand symmetry is fairly universal and has been documented in previous papers [99]. To address nucleic acid target symmetry, Macromodel 7.0 was used to flip and superimpose the docked pose on the crystallographic pose. This involves copying the complex consisting of the ligand docked to the target nucleic acid and then selecting to superimpose DNA bases from the copied structure onto the opposite DNA base of the original structure. Molecular superposition was performed using the "Superimpose Atoms" (SuprA) function followed by the "Rigid Superposition" (RigSA) function. In all cases, the resolution of the superposition was less than 0.15 Å. The superimposed structures were saved and the coordinates were used for RMSD calculations. Surflex docked poses are in a MOL2 file format which can be used directly by the Surflex program for RMSD calculations. Autodock docked poses are in a PDB file format and were converted to a MOL2 file format using Open Babel (2.1.1) [100] or iBabel [101] (2.0) prior to RMSD calculations. Docked poses of Autodock and Surflex in the target binding site were visualized using Autodock Tools.

Autodock and Surflex Scoring Function Methods. Rescoring of all top ranked Autodock and Surflex poses and the crystal structure poses was performed using the Autodock and Surflex scoring functions. To rescore all of the poses using the Autodock scoring function, the files were converted to Autodock PDBQT file format by merging all of the non-polar hydrogens. The Autodock *epdb* command was used to calculate a free energy of binding (kcal/mol) for each of the poses. The Surflex "score_list" command was used to rescore the top ranked poses using the Surflex scoring function. Macromodel was used to add hydrogens to the crystal structures and to the top ranked Autodock poses which normally only has polar hydrogens added for docking purposes. The Surflex scoring function ranks poses by an affinity score, pKd [97]. To fairly compare the docking poses for these two programs, the Surflex pKd results were converted to free energy of binding (kcal/mol), as previously described, where RT = 0.59 kcal/mol [102]:

$$\text{Free Energy of Binding} = RT \log_e(10^{-pKd}) \qquad (1)$$

Macromodel Energy of Binding Methods. Macromodel was used as a third, independent software to calculate the energy of binding of the poses using different force fields and solvation. All hydrogens were added, as previously described. The energy of binding was determined in structures with and without energy minimization of the hydrogens, as follows:

$$\text{Energy of Binding} = E_{\text{complex}} - E_{\text{ligand}} - E_{\text{nucleic acid}} \qquad (2)$$

Where: E_{complex} is the energy of the docked ligand in the target and the E_{ligand} and $E_{\text{nucleic acid}}$ represent the individually calculated energies. Energy minimization was performed by the Polak-Ribier Conjugate Gradient (PRCG) method for 1000 iterations with a convergence threshold of 0.05. The force fields were set at either Amber* or OPLS2005, with and without implicit water solvation to show the effects of these factors on the energy of binding. The experiments with no implicit water solvation were performed with distant dependent electrostatic treatment with a dielectric constant of 4.0 and an extended cutoff. The experiments with water solvation were performed with a constant dielectric electrostatic treatment with a dielectric constant of 1.0 and a normal cutoff.

Results and Discussion

Few studies have been performed to determine if molecular docking techniques such as Autodock and Surflex can dock ligands accurately to nucleic acids. We compare two poses derived from the docking calculations, the lowest RMSD pose for accuracy comparisons, and the top ranked pose for ranking comparison. A common metric for evaluation of accurate dockings is to calculate the RMSD between the crystallographic pose and the docked conformation. A level of significance of 2 Å will be evaluated to facilitate a comparison of these data to docking data in other reports [2, 86, 89, 99, 103]. When evaluating the ranking functions of the programs under different software conditions, only the single top ranking pose was used for comparing software conditions, as this is typically the mostly likely and facilitating pose that would be evaluated across large libraries of ligands that are used for virtual screening. The top pose was also inspected visually to determine the goodness of the ligand fit within the expected target site. Using these metrics, the optimal software conditions to maximize docking accuracy and ranking were "5 docks" and "2E7 energy evaluations" for Autodock and either the "Multistart 5" and "No Random" or the "Multistart 5" and "Random 5" for Surflex.

***Autodock 4 Docking Accuracy*.** Close examination of the dock with the lowest RMSD for each software parameterization shows that Autodock is able to accurately reproduce the crystal structure of several ligand-nucleic acid complexes to a resolution of less than 2 Å (Figure 5A). Taking ligand and nucleic acid target symmetry into account results in even lower RMSD poses for Pentamidine (ligand symmetry) and Ellipticine (nucleic acid target symmetry). Of the four ligands tested, Pentamidine is the only chemically symmetrical ligand. Accounting for this symmetry results in lower RMSD results since several of the poses that are docked in a flipped orientation relative to the crystal structure can be recalculated (Figure 5B). At first glance, the higher overall RMSD results for the optimal Ellipticine pose can be misleading as this appears to be a relatively poor docking. Visualization of the dockings reveals that the ligand is actually docked successfully into the intercalation site but lies in a flipped orientation rotated 180 degrees relative to the crystal pose. This flipped orientation of Ellipticine occurs for all of the lowest RMSD poses (Figure 5A) as well as the top ranked poses (Figure 6A). The

Figure 5. Autodock and Surflex accuracy: The dock with the lowest RMSD is presented, regardless of ranking. Figures A and C present the RMSD calculated without taking into account ligand or nucleic acid symmetry, for Autodock and Surflex, respectively. Figures B and D includes ligand and nucleic acid symmetry, for Autodock and Surflex, respectively. Black = Daunorubicin, Blue = Distamycin, Red = Ellipticine, Green = Pentamidine.

Figure 5. Autodock and Surflex accuracy: The dock with the lowest RMSD is presented, regardless of ranking.

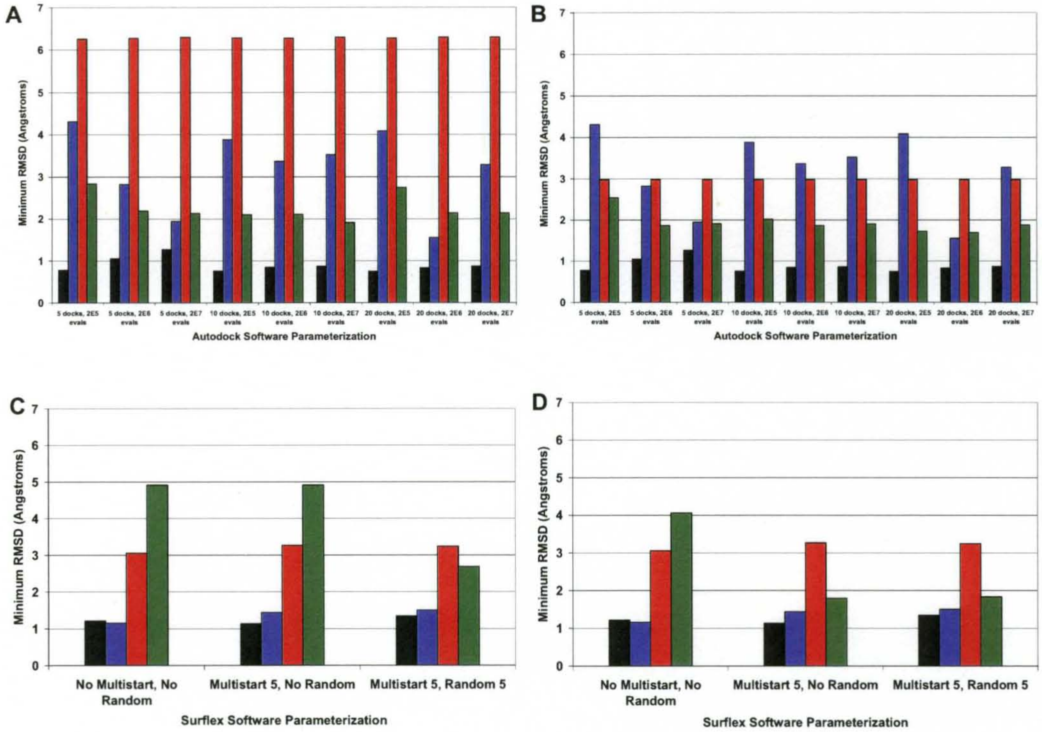
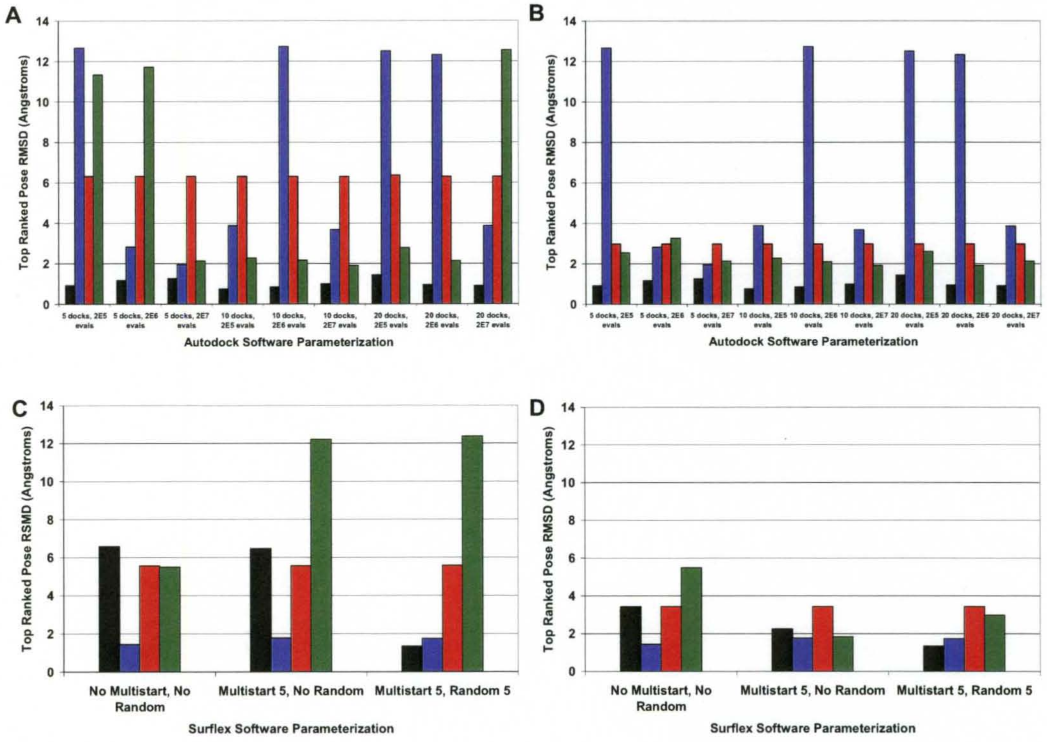


Figure 6. The top ranked pose by Autodock and Surflex. Figures A and C present the RMSD calculated without taking into account ligand or nucleic acid symmetry, for Autodock and Surflex, respectively. Figures B and D include ligand and nucleic acid symmetry, for Autodock and Surflex, respectively. Black = Daunorubicin, Blue = Distamycin, Red = Ellipticine, Green = Pentamidine.

Figure 6. The top ranked pose by Autodock and Surflex.



orientation and quality of the docked Ellipticine pose is partially explained by the symmetrical nature of the nucleic acid target, since Ellipticine can dock into the intercalation site not only from the orientation observed in the crystal structure but also from a flipped orientation with intercalation from the opposite side of the nucleic acid. Given that the Surflex RMSD calculator is based solely on the ligand poses and is irrespective of the nucleic acid target structure symmetry, the RMSD for Ellipticine is unusually high, even though Ellipticine is positioned well inside the intercalation site compared to the crystal structure. Thus, flipping and superposition of the docked pose on the crystallographic pose using Macromodel was necessary for an accurate comparison to the crystal structure. The fact that Ellipticine is docked in the intercalation site is encouraging, especially given the steric hindrance and tight fit typically associated with intercalation sites. Note that the Autodock grid is also large enough to allow for potential docking into the groove sites located near the intercalation site, so the intercalation dock is the preferred site. This emphasizes that RMSD values are only one metric for evaluating quality of docking poses and that the top poses should be visually inspected to check for ligand-target symmetry.

The lowest RMSD docking pose for Daunorubicin and Pentamidine are close to the resolution of the crystal structures, especially at the software conditions of "5 docks" and "2E7 energy evaluations". In particular, the RMSD for Daunorubicin is almost always lower than 1 Å. The RMSD values for Distamycin appear to be the most variable over the different software conditions, which is not surprising given that Distamycin has the highest number of rotatable bonds (14) compared to Daunorubicin (9), Pentamidine (12) and Ellipticine (0). The number of rotatable bonds for each molecule was defined by

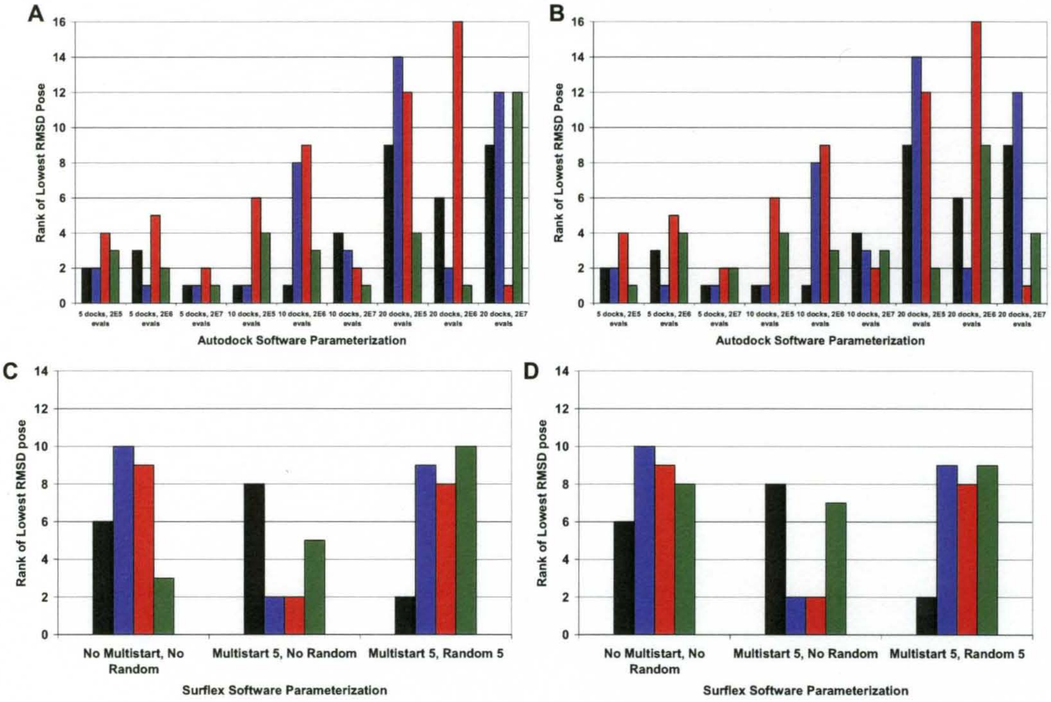
AutoDockTools using a united-atom representation that merges non-polar hydrogens [94]. AutoDockTools is used to automatically select the rigid "root" section of the ligand and the "branches" off of the "root" are subsequently defined as rotatable bonds [59]. Molecules with larger numbers of rotatable bonds are expected to take a larger number of energy evaluations to converge to an energy minimum due to a larger number of degrees of freedom and conformational states [59, 99]. The docking results for Distamycin are especially encouraging considering that most small molecules that are tested for therapeutic utility typically have less than 12 rotatable bonds [103]. The number of energy evaluations appears to be most important when the fewest number of docks (5) is used, and the accuracy of the Distamycin docking increases significantly with increasing number of energy evaluations. Moreover, once the number of energy evaluations used reaches 2E7, there appears to be no increase in docking accuracy when the number of docks is increased from 5 to 10 or 20. This finding is consistent with previous observations from ligand-protein studies that tested the effects of varying energy evaluations and number of dockings on docking accuracy [95]. Visualization of the Distamycin docking poses that have a resolution of greater than 2 Å show that even though the RMSD is higher than the cutoff, the ligand still occupies a similar space in the minor groove relative to the crystal structure. These results suggest that a software parameterization of "5 docks" combined with "2E7 energy evaluations" is acceptable, as the resolution of all of these docks with the exception of Ellipticine is less than 2 Å.

Autodock 4 Pose Ranking. The ability of Autodock to correctly rank the lowest RMSD docks must also be assessed as a particularly challenging aspect of molecular docking is scoring the docked poses correctly. The rank of the lowest RMSD pose out of all dockings for either Autodock and Surflex is shown in Figure 7. Autodock ranks the docked conformation by calculating a binding energy and sorting the results from lowest to highest energy. Ideally, the docked pose with the lowest binding energy would correspond to the docked pose with the lowest RMSD. In all software conditions, the top ranked dock for Daunorubicin achieves the RMSD cutoff of 2 Å. (Figure 6).

A number of poses with RMSD values less than 2 Å are produced for Distamycin and Pentamidine using several different software conditions. However, there are a number of top ranked poses for Distamycin, Ellipticine and Pentamidine in several software conditions that merit further discussion as these had higher RMSD values. It is critical to ascertain whether the high RMSD values associated with these poses is due to lack of consideration of either ligand or target symmetry or if the pose itself is of marginal quality. Visual inspection of the four top ranked poses for Distamycin with a resolution of greater than 12 Å RMSD suggests that the flipped orientation of the ligand relative to the crystal structure is the main cause of the high RMSD. However, the high RMSD cannot be ascribed solely to nucleic acid target symmetry, as the crystal structure shows that Distamycin is not centered around the minor groove and superposition of the docked pose results in poor visual overlap with the crystal structure. Instead there appears to be poor docking that is localized to the multiple terminal nitrogen groups, which float freely outside of the minor groove instead of the expected tight binding within the minor groove that is observed with the crystallographic structure. The marginal accuracy of these dockings may be influenced by the large number of rotatable bonds observed with Distamycin. This significantly increases the degrees of freedom and number of possible

Figure 7. The rank of the lowest RMSD pose out of all dockings. Figures A and C present the rank without taking into account ligand or nucleic acid symmetry, for Autodock and Surflex, respectively. Figures B and D includes ligand and nucleic acid symmetry, for Autodock and Surflex, respectively. Black = Daunorubicin, Blue = Distamycin, Red = Ellipticine, Green = Pentamidine.

Figure 7. The rank of the lowest RMSD pose out of all dockings.



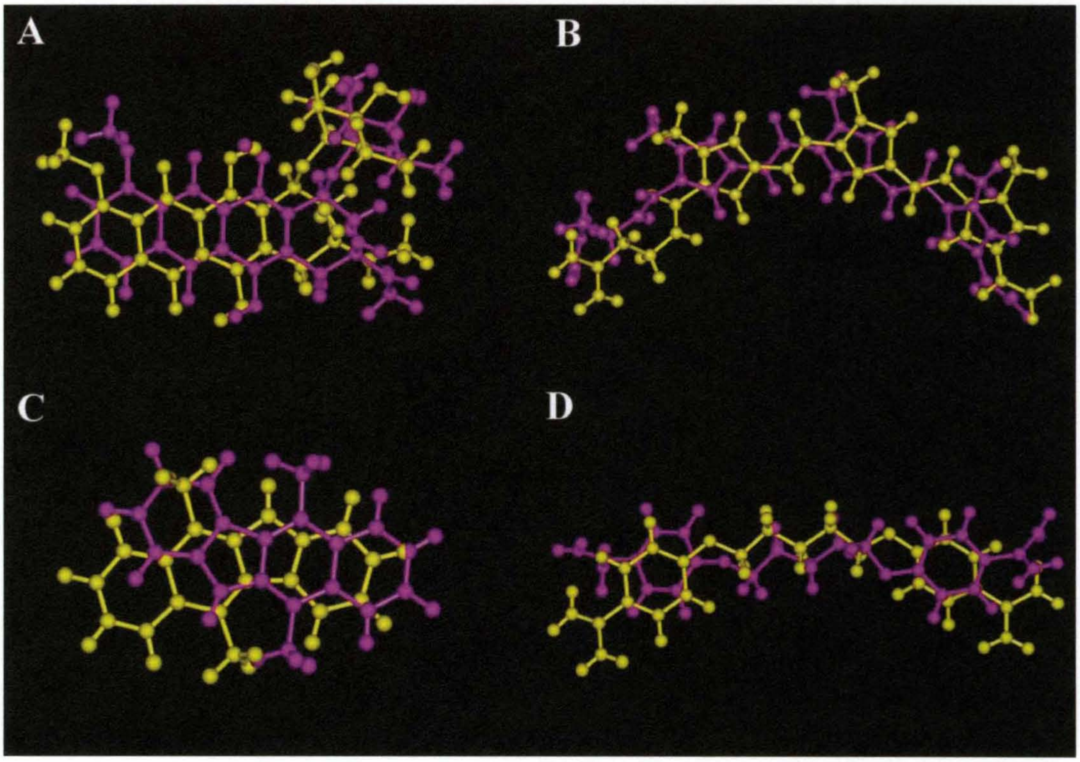
conformations of the ligand, making it challenging to dock to the target [59, 99]. With respect to Ellipticine, the high RMSD values appear to be due to a combination of the flipped orientation of the ligand which can be reassessed by accounting for nucleic acid target symmetry, and also by marginal overall alignment of the docked pose relative to the crystal structure. Pentamidine is a unique case where consideration of ligand symmetry into the RMSD calculations dramatically reduces the RMSD values for several top ranking poses (Figures 6B). This shows that the high RMSD is ascribed to ligand symmetry rather than to marginal docking quality and atom overlap.

In summary, there are several software conditions that appear promising with respect to ranking of the poses including "5 docks" with "2E7 energy evaluations" and "10 docks" with "2E5 energy evaluations". However, the real value in assessing Autodock performance lies in combining both docking accuracy and ranking of the docked results. A software parameterization of "5 docks" and "2E7 energy evaluations" appears best able to balance docking accuracy and ranking. By using this parameterization, docking of Daunorubicin, Distamycin and Pentamidine was achieved to a resolution of approximately 2 Å, while the intercalator Ellipticine was the most challenging dock, with a top pose resolution of approximately 3 Å. These docked conformations are also visually in close agreement with the observed crystal structure (Figure 8).

Surflex 2.11 Docking Accuracy. The Surflex docking results generally show that crystal structures are accurately reproduced (Figure 5). In all experiments, Daunorubicin and Distamycin are docked accurately to a resolution of less than 2 Å. Visualization of the lowest RMSD Ellipticine pose demonstrates that Ellipticine is docked in the correct

Figure 8. Comparison of the top ranked Autodock pose (magenta) to the PDB crystallographic pose (yellow) for the experiment with a software conditions of "5 docks" and "2E7 energy evaluations." (A) Daunorubicin, (B) Distamycin, (C) Ellipticine and (D) Pentamidine.

Figure 8. Comparison of the top ranked Autodock pose to the PDB crystallographic pose

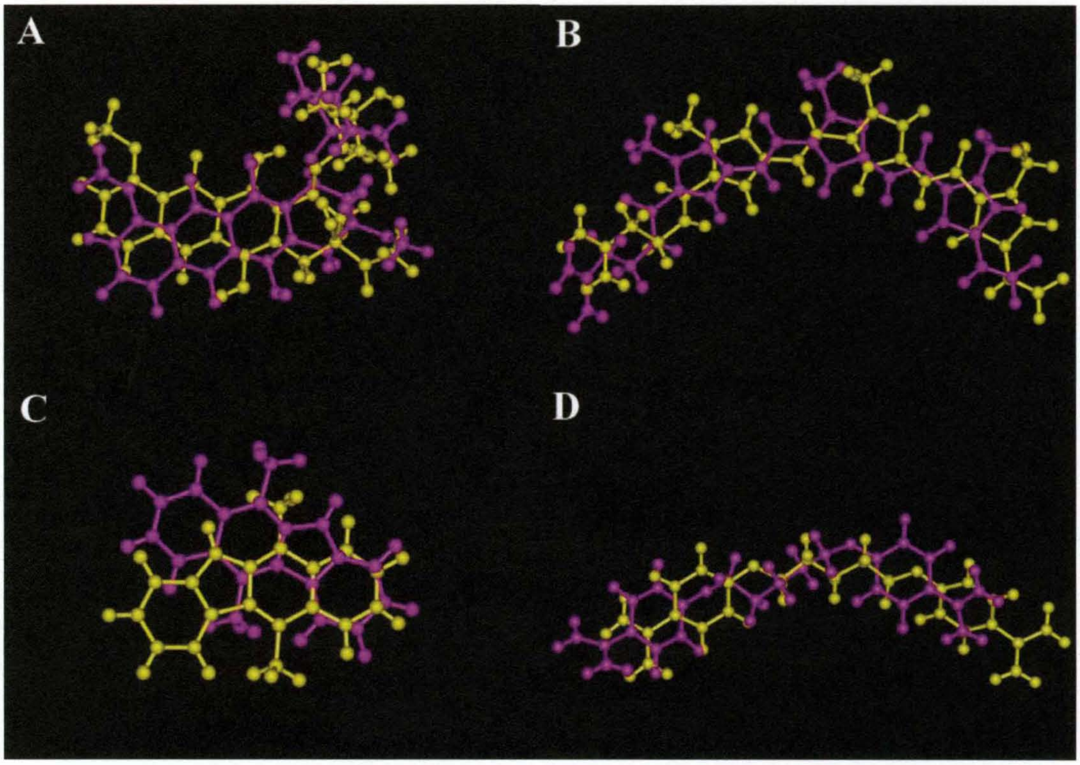


orientation relative to the crystal structure. The higher RMSD for the top Ellipticine pose relative to the other compounds appears to be due to the marginal alignment of the ligand structure with the crystal structure. A similar marginal overlap was observed for the Autodock Ellipticine poses, as described previously. Importantly, Ellipticine is located well inside the intercalation site. For Pentamidine, incorporation of ligand symmetry into the RMSD calculation results in significant increases in the docking accuracy for all software conditions, with the lowest RMSD structures occurring with the "Multistart 5" only experiment, and the "Multistart 5" and "Random 5" combination experiment. This is attributed to inclusion of poses that were docked in a flipped orientation that initially had RMSD values greater than 12 Å, but subsequently have significantly lower RMSD values after taking into account ligand symmetry. With respect to docking accuracy, addition of the "Multistart 5" option produces a better docked pose for Pentamidine. This supports the hypothesis that initiating the docking of the ligand from multiple points surrounding the nucleic acid target increases the accuracy of the dockings. Interestingly, the addition of the "Random 5" option in combination with the "Multistart 5" option did not significantly impact the lowest RMSD dock produced for these test ligands. The "Random 5" option generates 5 randomized X,Y,Z coordinate positions of the atoms at the initial starting position of the ligand [97]. Most importantly, Surflex is able to dock the ligands to the nucleic acid targets and produce docking results with RMSD values close to the resolution of the observed crystal structure.

Surflex 2.11 Pose Ranking. Ranking of Surflex results is performed by maximizing the Surflex Overall Score, which consists of an affinity score of the ligand for the target. Ideally, a maximal Surflex Overall Score would correspond with the lowest RMSD pose.

Figure 9. Comparison of the top ranked Surflex pose (magenta) to the PDB crystallographic pose (yellow) for the experiment with a software parameterization of "Multistart 5" and "Random 5." (A) Daunorubicin, (B) Distamycin, (C) Ellipticine and (D) Pentamidine.

Figure 9. Comparison of the top ranked Surflex pose (magenta) to the PDB crystallographic pose (yellow) for the experiment with a software parameterization of "Multistart 5" and "Random 5."



Inspection of the RMSD values for the top Surflex docks ranked by maximal Surflex Overall Score are at first glance misleading (Figure 7). In particular, the experiment that included the "Multistart 5" and "No Random" options and the experiment with the "Multistart 5" and "Random 5" options initially appear to have a poor docking pose for Pentamidine. However, closer visual inspection of the docked conformation relative to the crystal structure pose again emphasizes the use of symmetry for RMSD calculations where appropriate (Figure 9), which reduces the RMSD to under 2 Å.

For all of the experiments, Ellipticine is docked in a flipped orientation in the intercalation site, which was initially thought to be the major factor influencing the high calculated RMSD value. However, even after accounting for nucleic acid target symmetry, Ellipticine has still only minimal overlap with the crystal structure. Inspection of the top ranked dock for Daunorubicin for the software parameterization with "No Multistart" and "No Random" and the software parameterization with "Multistart 5" and "No Random" options appears to show Daunorubicin in a flipped orientation relative to the crystal structure. The Daunosamine ring occupies the minor groove, which is similar to the ring location in the crystal structure. After taking into account the nucleic acid target symmetry, the docking pose RMSD values for both the "Multistart 5" and "No Random" experiment and the "Multistart 5" and "Random 5" experiment are dramatically improved, to a resolution of 3.4 Å and 2.3 Å, respectively. Finally, all Surflex software conditions docked Distamycin to the target at a resolution of less than 2 Å. These results emphasize the importance of not only calculating RMSD values for docked poses, but also visualizing results to check for reasonable docking conformations.

The software parameterization with "Multistart 5" alone and the software parameterization with both "Multistart 5" and "Random 5" appear to produce the top ranked results with the lowest RMSD structures, compared to the software parameterization of "No Multistart" and "No Random" options. The top ranked pose for Daunorubicin using the "Multistart 5" and "Random 5" software parameterization has an RMSD of 1.3 Å and is superior to the top ranked dock for the other Surflex experiments. Both software conditions dock Distamycin and Ellipticine comparably with respect to the RMSD of the top ranked Surflex pose. For Pentamidine, the top ranked pose for the "Multistart 5" and "No Random" option experiment has a marginally better RMSD for the top pose compared to the top pose from the "Multistart 5" and "Random 5" experiment. Overall, the performance of the "Multistart 5" and "No Random" experiment and the "Multistart 5" and "Random 5" experiments are comparable.

Extended Parameter Optimization for Autodock and Surflex. While the overall docking results for Surflex and Autodock generally show the ability to accurately reproduce the crystal structure and rank the results, it is important to determine the reason for some of the more challenging dockings such as Ellipticine and Distamycin. One possibility for the marginal docking accuracy could be an inadequate number of iterations (number of docks and energy evaluations for Autodock, and multistart number and random parameters for Surflex) of the software. If this is the case, it would be expected that increased docking accuracy and ranking could be obtained by increasing the exploration of the Autodock and Surflex parameters.

To investigate this possibility for Autodock, the docking experiments with the four ligands were repeated after increasing the number of dockings from 5 to 50 and the number of energy evaluations from 2E7 to 5E7. The number of dockings were selected based on previous applications of the software [59]. The number of energy evaluations was increased to 5E7, which is consistent with the number of energy evaluations used in previous protein docking experiments [95]. A similar approach was taken with Surflex by increasing the Multistart parameter from 5 to 10 and the Random parameter from 5 to 10. However, Jain *et al.* had previously seen only marginal improvement in increasing the Multistart parameter greater than 5 with protein docking [97].

Evaluation of the docking accuracy (Figure 10) and ranking (Figure 11) results show that there is no benefit in docking accuracy or ranking for either Autodock or Surflex by extending dockings and evaluations of software parameters. Moreover, the Autodock experiments took approximately 25 fold longer under conditions of 50 docks and 5E7 energy evaluations compared to conditions of 5 docks and 2E7 energy evaluations. Surflex took approximately 5 times longer under conditions of Multistart 10 and Random 10 compared to Multistart 5 and Random 5. Even if the extended experiments showed improved docking accuracy and ranking, the increase in computational time could be a limiting factor for use in virtual screening applications. In summary, the results suggest that the originally optimized Autodock conditions of 5 docks and 2E7 energy evaluations and Surflex conditions of Multistart 5 and Random 5 are optimized for molecular docking to nucleic acids.

Evaluation of the Autodock and Surflex Scoring Functions. As the docking accuracy and ranking does not appear to be related to suboptimal software parameterization, another possible contribution to marginal docking may be from the

Figure 10. Autodock and Surflex Parameterization Accuracy. The dock with the lowest RMSD is presented, regardless of ranking. Figures A and C present the RMSD calculated without taking into account ligand or nucleic acid symmetry, for Autodock and Surflex, respectively. Figures B and D includes ligand and nucleic acid symmetry, for Autodock and Surflex, respectively. Black = Daunorubicin, Blue = Distamycin, Red = Ellipticine, Green = Pentamidine.

Figure 10. Autodock and Surflex Parameterization Accuracy.

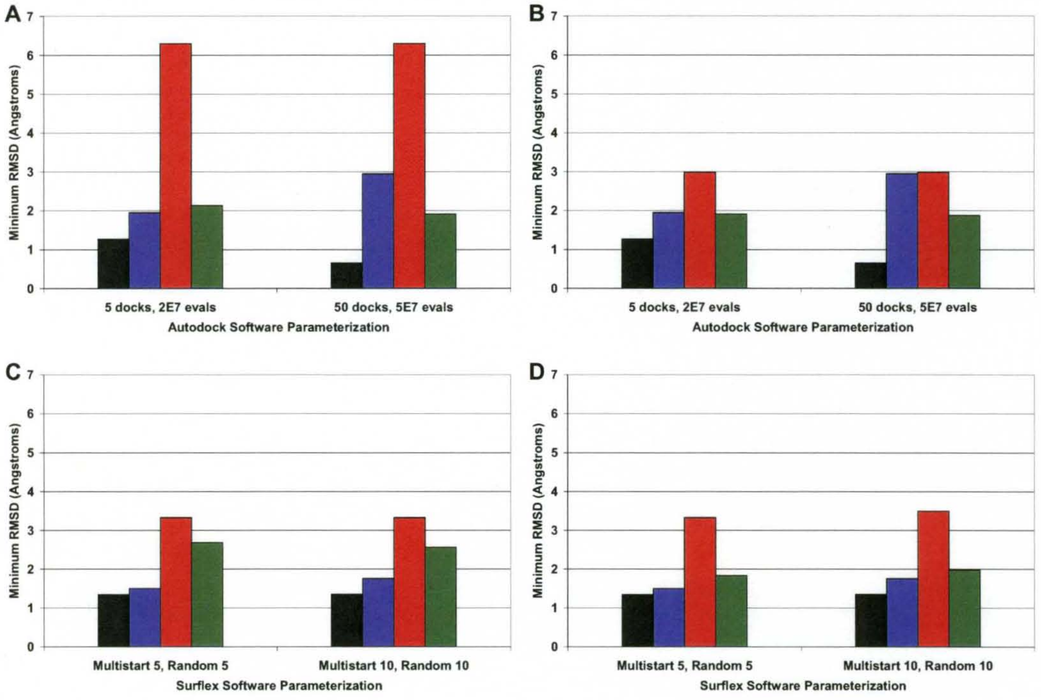
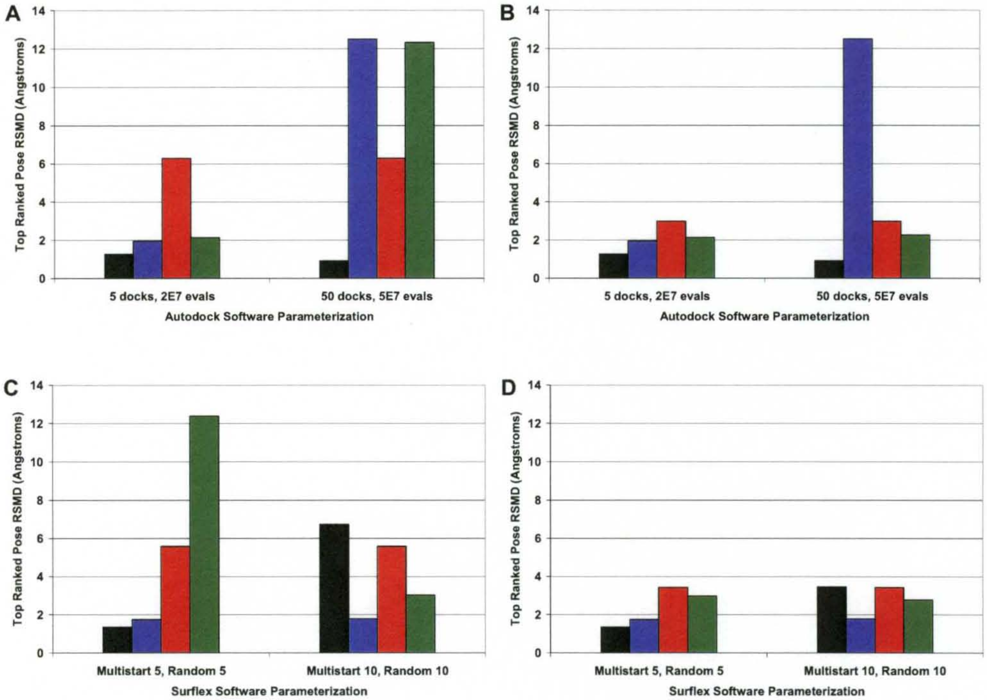


Figure 11. Autodock and Surflex Parameterization Ranking. Figures A and C present the RMSD calculated without taking into account ligand or nucleic acid symmetry, for Autodock and Surflex, respectively. Figures B and D include ligand and nucleic acid symmetry, for Autodock and Surflex, respectively. Black = Daunorubicin, Blue = Distamycin, Red = Ellipticine, Green = Pentamidine.

Figure 11. Autodock and Surflex Parameterization Ranking.

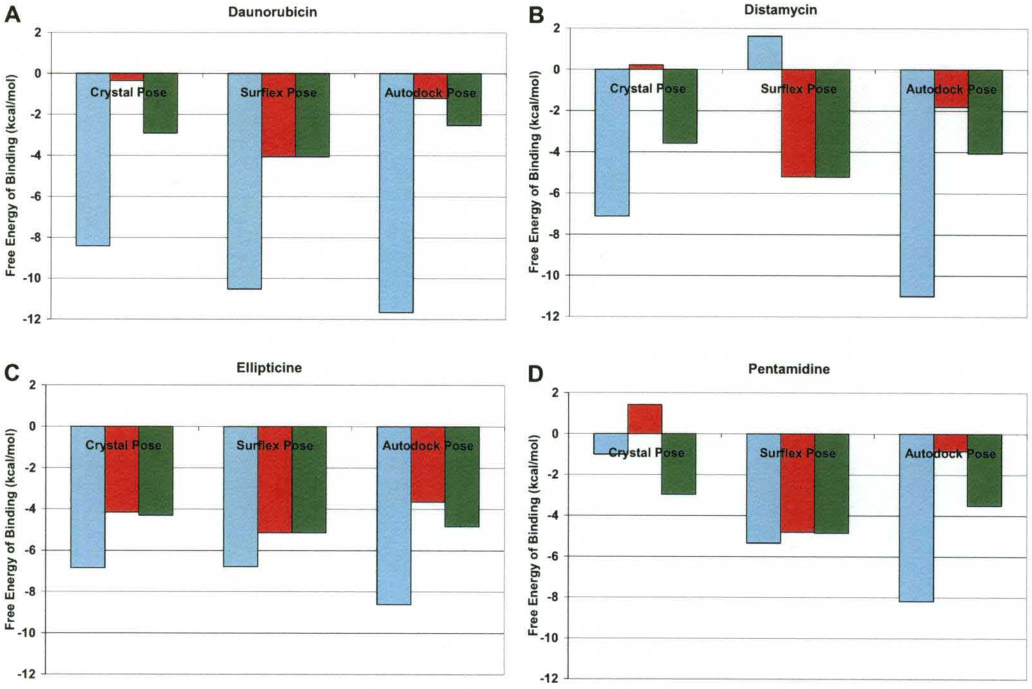


scoring functions of these programs. This is possible given that scoring functions are one of the major challenges of current docking programs [104]. To investigate this possibility, the crystal structure and the top ranked poses for each method were rescored using both the Surflex and Autodock scoring functions. The poses were scored and ranked according to the lowest free energy of binding. An additional molecular mechanics method was selected to calculate the energy of binding of the crystal pose, Autodock, and Surflex poses. This was useful as the added hydrogens could also be selectively energetically minimized, which highlighted the hydrogen atom treatment as a potential pitfall. Macromodel 9.5 was used to determine the effects on the energy of binding of using either the OPLS2005 or Amber* force field with and without water as an implicit solvent. These experiments investigated if the limitations in the ranking of the software were related to the scoring functions for these programs.

Scoring of Poses by Autodock and Surflex. The direct comparison of the Surflex and Autodock Scoring Functions is shown in Figure 12. Unsurprisingly, the Surflex scoring function tends to score the Surflex poses the best while the Autodock scoring function tends to score the Autodock poses the best. The Surflex scoring function scores the Autodock poses reasonably well, with a low free energy of binding. In general, the Autodock scoring function produces results with the lowest free energy of binding. Both Autodock and Surflex appear to typically score either the Autodock or Surflex poses as having lower free energy of binding compared to the crystal pose. The Surflex scoring function produces a "Static" score (red, Figure 12) and an "Optimized" score (green Figure 12) when scoring an individual pose. The "Static" score applies the Surflex

Figure 12. Comparison of the Free Energy of Binding for the Crystal pose, Autodock top ranked pose and Surflex top ranked pose for various ligands using the Autodock and Surflex Scoring Functions. Blue = Autodock Scoring Function. Red = Surflex Static Scoring Function. Green = Surflex Optimized Scoring Function. (A) Daunorubicin, (B) Distamycin, (C) Ellipticine and (D) Pentamidine.

Figure 12. Comparison of the Free Energy of Binding for various ligands.

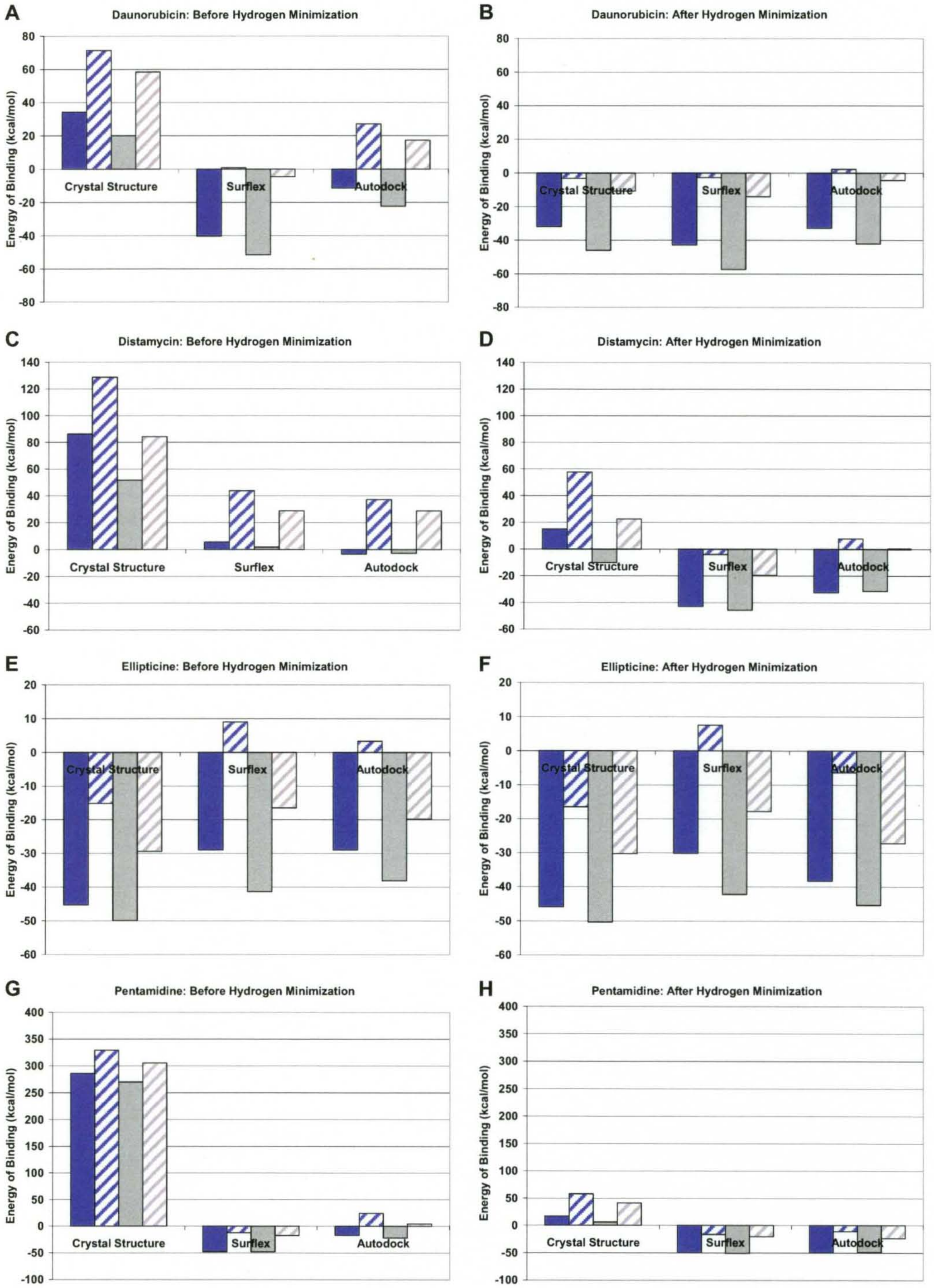


scoring function directly to the input pose, with no energy minimization. The "Optimized" score performs a gradient energy minimization and subsequently scores the pose. Scoring of the Surflex poses using the Surflex scoring function reveals little difference between the Static score and Optimized score. On the other hand, Autodock and the crystal structure scores are significantly improved when comparing the "Static" score to the "Optimized" score. One possible explanation for this difference is how the hydrogens are accounted for by these docking programs.

Hydrogen Atom Treatment of Poses Can Significantly Effect Free Energy of Binding. It appears from Figure 12 that the significant difference in the "Static" and "Optimized" Surflex scoring function scores for Autodock and the crystal poses could be influenced by the way hydrogen atoms are added to these structures. In order to determine if this is the case, it is important to first address the way hydrogens are normally accounted for by these programs. Surflex adds all hydrogens on the ligand prior to docking so all of the hydrogens are present during scoring. The crystal structure does not have any hydrogens added. Autodock uses a United Atom force field which takes into account "polar" hydrogens that are attached to electronegative atoms [59]. “Non-polar" hydrogens attached to carbon atoms are merged and the charge is added to the nearby carbon atom [85]. To evaluate whether the trends in Figure 12 could be influenced by the way hydrogens are handled by the docking programs, Macromodel was used to add all hydrogens to the ligands and their binding energies were recalculated both before and after energy minimization of the hydrogens (Figure 13). Comparing the binding energy of the poses before and after minimization of the hydrogens shows that the most significant decrease in energy after minimization is seen with the crystal structure.

Figure 13. Calculated Energy of Binding by Macromodel for the Crystal pose, Autodock top ranked pose and Surflex top ranked pose for various ligands using the Amber* and OPLS2005 Force Fields with and without implicit water solvation. Solid Blue =OPLS2005, no implicit water solvation. Blue with Hatches = OPLS2005, with implicit water solvation. Solid Gray = Amber*, no implicit water solvation. Gray with Hatches = Amber*, with implicit water solvation. (A) and (B): Daunorubicin, before and after hydrogen minimization, respectively. (C) and (D): Distamycin, before and after hydrogen minimization, respectively. (E) and (F): Ellipticine, before and after hydrogen minimization, respectively. (G) and (H): Pentamidine, before and after hydrogen minimization, respectively.

Figure 13. Calculated Energy of Binding by Macromodel for the various ligands.



However, there is also a substantial reduction in the energy of binding for Autodock. The Surflex binding energies appear to be the least affected presumably because all hydrogens were accounted for during docking and scoring. The results in Figure 13 are important because a molecular mechanics approach was used to assess each of the poses for the docking programs with two force fields and two solvation approaches. These results show that Surflex appears to consistently produce the docked poses with the lowest energy of binding. This suggests the hydrogen atom treatment is an important consideration when scoring docked poses and can substantially influence scoring and energy calculations. It is interesting to note that Ellipticine, which has the fewest number of rotatable bonds and hydrogen atoms is least effected by hydrogen atom treatment.

Effects of Force Field Choice and Solvation on Energy of Binding. A series of experiments was performed to test the effects of using either the Amber* and OPLS2005 force fields, with and without implicit water solvation, on the energy minimization and the calculated energy of binding of the ligands to the targets. The Amber* force field was selected because the Autodock force field is parameterized based on the Amber force field [59, 85]. OPLS2005 was chosen because it is an updated general force field from the original OPLSAA force field that has demonstrated utility in evaluating protein structures [105]. The calculated energy of binding of the top ranked ligand poses, before and after energy minimization of the added hydrogens are shown in Figure 13. The force field choice and solvation effects can substantially influence the calculated energy of binding. For the structures where the hydrogens were energetically minimized, the use of the Amber* force field with inclusion of water solvation appears to produce energy of binding results that are most consistent with the results in Figure 12 that were obtained using the Autodock and Surflex scoring functions. The energy of binding of the Autodock and Surflex poses appears substantially lower than the crystal structures, with the exception of Ellipticine. It appears in these cases that for the Autodock and Surflex poses, the addition of implicit solvation in just the energy minimization is not advantageous and not indicative of a favorable binding event. In total, this shows that force field selection and solvation factors can contribute substantially to scoring and ranking docked poses and this could be one of the main challenge of docking ligands to nucleic acids.

Crystal Structure Energies are not necessarily the "Minima". The free energy of binding for the crystal structure and top ranked Autodock and Surflex poses, determined by either the Autodock or Surflex scoring function is shown in Figure 12. The top ranked Autodock or Surflex pose almost universally has a comparable or lower free energy of binding compared to the reference crystal structure. This is true irrespective of the scoring function. These results are supported by the molecular mechanics results in Figure 13, where the calculated energy of binding for all ligands, apart from Ellipticine, is comparable to or lower than the crystal structure. These results are important for several reasons. First, the crystal structures should not be assumed to be the energetically minimized conformation in the nucleic acid target, as the structure is a product of experimental data and the original force field it is fitted to. Interestingly, the energy of the lower resolution crystals, Distamycin (2.2 Å) (Figure 13D) and Pentamidine (2.1 Å) (Figure 13H), appear to have more variability between the energy of the top ranked poses and the crystal structure compared to the higher resolution crystal structures Daunorubicin (1.4 Å) (Figure 13B) and Ellipticine (1.5 Å) (Figure 13F). This suggests that the quality and resolution of the crystal structure may be a consideration when performing docking studies and evaluating poses. However, it is also a function of flexibility of the ligand as Distamycin and Pentamidine are the most flexible. Another reason these results are important is that the docked poses such as Distamycin that initially appeared to be of only marginal accuracy by RMSD compared to the crystal structure are better than initially thought with respect to the energy of binding, which implies that the crystal structure ligand pose may not be optimal to start with.

Overall Comparison of Autodock and Surflex Performance. In assessing the overall performance of Autodock and Surflex, several facets of docking must be compared including docking accuracy, docking ranking, computational speed, and even ease of use. Both Autodock and Surflex have comparable performance in accurately reproducing the crystal structure and ranking the poses, particularly with software conditions of "5 docks" with "2E7 energy evaluations" and "Multistart 5" and "Random 5" respectively. However, one important factor where performance differs substantially are the computational resources required for docking. Using 2.0 GHz AMD Opteron 246 processors, Surflex performed the dockings significantly faster than Autodock for all ligands tested. The average time to complete each Surflex docking with a software parameterization of "Multistart 5" and "Random 5" was just under 8 minutes while Autodock with a software parameterization of "5 docks" with "2E7 energy evaluations" took approximately 76 minutes. Given that the docking accuracy and ranking results were comparable, the significantly faster docking speed of Surflex makes it particularly well suited for virtual screening applications where large numbers of ligands are screened. Surflex is also superior with ease of use, as it is a single executable application with direct input from a MOL2 file format. Autodock requires file conversion from a MOL2 into a PDBQT file format prior to performing molecular dockings. For these reasons, under the tested software conditions, we show Surflex is a superior software package for virtual screening of nucleic acids in the system reported here.

Comparison of Results to Previous Studies. Relatively few molecular docking studies have been performed with nucleic acids. In comparing the data presented in this paper to other docking papers, we placed particular emphasis on the evaluation of the accuracy of the top ranked pose returned by either Surflex or Autodock. This is a logical approach for assessing docking software performance for virtual screening applications, since when screening a large ligand database, only the evaluation of the top ranked pose may be computationally feasible. Several previous studies have focused on utilization of the DOCK program for molecular docking of ligands to nucleic acids. Grootenhuis *et al.* used DOCK to target the minor groove, major groove and an intercalation site on duplex DNA while more recently, Chen *et al.* successfully targeted the major groove of RNA [33, 36-37]. Yan *et al.* targeted an RNA tetraloop structure and demonstrated docking at a similar resolution to what was observed in our study of docking ligands to DNA targets [72]. Rohs *et al.* recently developed a molecular docking approach utilizing a Monte Carlo algorithm that successfully demonstrated the binding of methylene blue to DNA by minor groove and intercalation binding modes [38]. However, methylene blue has only four rotatable methyl groups with fewer degrees of freedom than several of the more conformationally complex ligands tested in this study [38].

One report of docking studies to nucleic acids using Autodock was performed by Evans *et al.*, who demonstrated the ability of a previous version of Autodock to accurately predict binding of minor groove binders to their respective nucleic acid targets [34]. A direct comparison of all of the results from the Evans paper and this study is difficult due to different operating conditions and software versions for Autodock; however, some differences are noteworthy. One limitation of the previous study is that while the number of energy evaluations was varied, the maximum number of evaluations performed was only 2.5E6. Based on our studies, we found that 2E7 energy evaluations was optimal for docking accuracy and pose ranking. Another consideration is that in this previous study the number of dockings was kept constant. We evaluated the parameters by varying both the number of docks and energy evaluations to determine which combination of software parameters is best for virtual screening applications. Similar to the results in this paper, Evans did find that in general, increasing the number of energy evaluations increased the accuracy of the predicted pose, with respect to the crystal structure [34]. However, we also found that using fewer numbers of dockings while concurrently increasing the number of energy evaluations increases both pose accuracy and ranking. This is presumably due to a more complete exploration of the energetic landscape surrounding the ligand-target interaction. This has important implications for virtual screening where of crucial importance is the accuracy of the top ranked pose. Generally, the results of Evans *et al.* are consistent with results in this paper, and show that Autodock is able to successfully predict the binding of multiple minor groove binders to their targets at a resolution of approximately 2 Å [34]. However, based on the data herein, we recommend using more energy evaluations and fewer numbers of docks for virtual screening applications to produce the best top ranked dock. While the results in this paper expand and add value to previous Autodock work targeting nucleic acids, importantly, we show that the results with Surflex in particular are very useful, applicable, and the first published study to demonstrate successful molecular docking of intercalators or minor groove binders to nucleic acid targets using this software.

Conclusions

The results reported here support the primary objective of this work, which is to test Autodock 4 and Surflex 2.11 for accurately reproducing ligand-bound nucleic acid structures. This is a critical first step in validating these software for future use in targeting specific nucleic acid structures. Even given the aforementioned limitations and uncertainties of using Autodock 4 and Surflex 2.11 with nucleic acids, these results show that these software can accurately reproduce the crystal structures of both groove binders and intercalators. Ours is one of only a few studies to date to have shown that nucleic acids can be successfully targeted using these docking methods. Our results show that an Autodock 4 software condition of "5 docks" and "2E7 energy evaluations" is the best for combined docking accuracy and ranking. The Surflex 2.11 software conditions of "Multistart 5" and "No Random" and "Multistart 5" and "Random 5" appear equally good at producing top ranked structures with low RMSD values relative to the crystal structure. Extended experiments testing further increases in Autodock and Surflex parameterization did not improve docking accuracy or ranking. The most challenging ligand to dock accurately was Ellipticine, which was no surprise given the small pocket in the nucleic acid and tight fit associated with the binding of ligands into the intercalation site. Given that the Autodock and Surflex scoring functions for ranking the docked poses were parameterized based on protein-ligand structures, the ranking results are particularly encouraging [2, 86]. Both programs are able to return a top ranked pose with approximately 2 Å RMSD for Daunorubicin, Distamycin and Pentamidine and a pose with approximately 3 Å RMSD for Ellipticine. It is important to consider that while the docking accuracy and pose ranking of these programs is comparable, Surflex performs docking much faster than Autodock under the optimized software conditions in this paper. Surflex also requires less manipulation of input files, suggesting that Surflex is preferred for virtual screening applications for systems similar to presented here.

Based on these docking studies, several points should be strongly considered when performing molecular docking with nucleic acids and evaluating docked poses. Docking parameters should be explored in detail since suboptimal software conditions can significantly impact the accuracy and ranking of the docked poses. When evaluating docked poses, visualization of the most promising docking poses should be performed as well as calculation of RMSD values. It is crucial to also account for both ligand and target symmetry by either including ligand symmetry in RMSD calculations or performing molecular superposition to account for nucleic acid target symmetry. Given the conformation and structural heterogeneity of proteins, target symmetry is less likely with respect to docking. However, nucleic acid targets are much more likely to exhibit symmetry due to the simple base pair composition and the nature and geometry of the nucleic acid strand associations. Another consideration when performing docking is the hydrogen atom treatment of the software, as this can significantly impact the free energy of binding. These studies also demonstrated that force field and solvation selection can dramatically effect the binding energy. Finally, selection of high quality and high resolution crystal structures is especially important when using these structures as reference conformations to evaluate docking poses. Based on the results in this paper, it is important to consider that the crystal structure does not necessarily represent the energetically minimized pose with respect to the poses generated by docking software. These findings have important implications not only in the field of chemistry and computational biology, but also in the area of organic small molecule synthesis using structure-based drug design. Many previous efforts at rational drug design have focused on time-consuming and expensive small molecule synthesis methods. If reliable, molecular docking allows for the construction of virtual libraries of molecules that can be docked against any nucleic acid target of interest. One of the logical next steps in molecular docking to nucleic acids is the development of rules to select ligands that may bind nucleic acid targets with affinity and specificity. These experiments suggest that molecular docking techniques may have particular value as a virtual screening precursor step to full chemical synthesis of drug candidates.

CHAPTER III

DEVELOPMENT OF *IN SILICO* PREDICTIVE METRICS THAT GOVERN SMALL MOLECULE – NUCLEIC ACID INTERACTIONS

Work in our previous chapter described the validation of the molecular docking software Surflex and Autodock for the purposes of reproducing the crystal structure of the minor groove binder and intercalator small molecules bound to nucleic acid targets. The results were significant as they demonstrated that these software can be used for molecular docking to nucleic acids. However, this work involved rationalization of known crystal structure data by docking the small molecules to a single nucleic acid target. The question remains whether a compound can be screened against multiple nucleic acids *in silico* for the purposes of predicting binding mechanism and sequence and structural selectivity. Determining if predictive *in silico* metrics can be developed to answer this question is the focus of this chapter.

This chapter details a novel approach to predict the binding mechanism and sequence and structural selectivity of small molecules for nucleic acids. We describe the construction of an *in silico* nucleic acid library and the docking of the small molecules from Chapter II (daunorubicin, distamycin, ellipticine and pentamidine) to the array of nucleic acids. Metrics were developed that successfully classify these compounds as either groove binders and intercalators, with moderate success at predicting sequence and structural selectivity. The metrics were further tested on several new triplex and quadruplex binding small molecules that our lab has recently discovered. Finally, using the *in silico* metrics, an extensive 67 member small molecule library for which *in vitro* nucleic acid sequence and structural binding data exists, was classified on the basis of binding mechanism and sequence selectivity. This was the most robust test of the metrics as the compound library was highly heterogeneous with respect to binding mechanism of action and sequence preference. In total, we demonstrate that the metrics as described here can generally successfully predict the mechanism of binding of a ligand to a nucleic acid *in silico* although it was generally more challenging to predict sequence and structural selectivity. A summary comparison of the performance and limitations of Surflex and Autodock is also detailed. The new information described here can facilitate large scale virtual screening efforts that can be used to discover new small molecules *in silico* that bind to a specific site on a nucleic acid structure and with a desired binding mechanism.

DEVELOPMENT OF *IN SILICO* PREDICTIVE METRICS THAT GOVERN SMALL MOLECULE – NUCLEIC ACID INTERACTIONS

Patrick A. Holt, Jonathan B. Chaires, John O. Trent.

Introduction

Knowledge about the structure and function of nucleic acids has increased dramatically in recent years. It is now known that nucleic acids are highly polymorphic and can adopt physiologically relevant structures *in vivo* that are promising targets for drug development. There is increasing evidence suggesting that DNA is altered in many neoplastic conditions and there are many nucleic acid structures that are intriguing targets for small molecule drug discovery [106]. G-quadruplex structures are but one example. These structures are found in increased prevalence in the single stranded telomeric ends of chromosomes. Stabilization of G-quadruplexes through small molecule targeting has been shown to inhibit cancer cell life by inhibiting telomerase association with the chromosome. As telomerase is overexpressed by cancer cells and not normal cells, this is a potentially effective strategy for selectively targeting tumor cells [53-55].

There are multiple known small molecules that bind to G-quadruplexes. One example is the porphyrin TmPyP4 which has potential anti-cancer properties *in vivo* [107-108]. Unfortunately, many of these molecules, TmPyP4 included, suffer from poor selectivity and are known to bind to many other nucleic acid structures and sequences *in vitro*, which is a major concern for further clinical development [10, 109]. This poor selectivity may be in large part because many small molecules are designed or synthesized only considering the binding to a single target of interest and may not take into account binding and interactions with other potential targets *in vitro* and *in vivo*. There is a critical need to determine if the mechanism of action and sequence and structural selectivity of small molecules for nucleic acids can be predicted *in silico*. This would allow for the virtual screening of millions of molecules *in silico* for the best "hit." We describe here a novel computational strategy here to address this unmet need.

Before detailing our *in silico* approach for predicting small molecule-nucleic acid interactions, we briefly review the important binding mechanisms of compounds to nucleic acids. Small molecules can interact with nucleic acids by two main modes of binding; groove binding and intercalation [110]. The minor groove of DNA provides a site for many small molecules to bind because of the favorable geometry of the groove and because there is less competition from polymerases and proteins that typically target the major groove of duplex DNA [5]. Small molecules that bind to the minor groove typically have intrinsic curvature present or have the capability of existing as a stable, low energy conformation that is compatible with the geometry of the minor groove [111]. This compound curvature or "crescent shape" is an important property of many minor groove binding small molecules as this allows the compound to bind between the walls of the minor groove [111-112]. However, compound curvature is not an absolute requirement for compounds that bind to the minor groove, as some linear diamidine compounds have been identified that possess minor groove binding activity [113]. A few examples of small molecules that bind to the minor groove are DAPI, pentamidine and distamycin.

The second main binding mechanism of small molecules to nucleic acids is intercalation which occurs by insertion of the molecule between adjacent base pairs in the nucleic acid. Intercalation of a small molecule into DNA exerts a profound change on the structure of DNA. In order for the intercalator to stack between the adjacent bases, the nucleic acid must unwind partially and increase in length [110]. "Classical" intercalators typically possess a fused, planar aromatic ring system which allows a small molecule to insert between adjacent base pairs. In almost all cases, the small molecules also possess a cationic external charge and many times bind cooperatively to DNA [110]. Molecules that intercalate into DNA include ethidium bromide and acridine based molecules.

Compounds can also interact by hybrid methods where contributions of both intercalation and groove binding are involved. Additionally, in some cases, molecules can stack onto the ends of specific DNA structures ("end-stacking") such as G-quadruplexes, which involves interaction of the ligand with the guanine quartet on the one side and the flanking DNA loops on the other side. This is in contrast to intercalation into G-quadruplexes which occurs when a compound inserts between two adjacent guanine tetrads. Finally, molecules can have chemical properties that allow intercalation of part of the molecule with concomitant groove binding of substituents which is referred to as "threading" intercalation. Generally, however, small molecules are divided along the main categories of minor groove binding and intercalation.

We focus here on determining if predictive *in silico* rules can be developed to predict the nucleic acid binding mechanism and sequence specificity of a small molecule *in silico*. This would provide valuable information as the rules could be applied to virtually screen large numbers of small molecules to identify ones that bind with a known mechanism of action to a specific nucleic acid target. Large scale *in silico* molecular docking of small molecules to a target of interest is becoming an accepted approach for discovering novel small molecules for drug development. This field has largely focused on proteins until recently as many of the software have been designed with proteins in mind. However, validation of various software with nucleic acids has been successful and recent evidence suggests that the molecular docking software, Surflex and Autodock has particular utility for the virtual screening of large numbers of compounds to promising nucleic acid targets [34, 63]. We previously reported (as described in Chapter II) the use of Surflex and Autodock to successfully reproduce the known crystal structure pose of a collection of small molecules that bind by groove binding and intercalation to nucleic acid targets [63]. While these studies were successful, they relied on the presence of a single *in silico* structure of a small molecule with nucleic acid and did not determine if the docking software can predict sequence or structural selectivity. The question remains whether rules can be developed for Surflex and Autodock that can predict whether small molecules will groove bind or intercalate and to which sequences and DNA morphologies that the small molecules prefer.

We report here the development of *in silico* rules that can be used to predict the mechanism of action and the sequence specificity of an array of small molecules. An initial set of four small molecules (daunorubicin, distamycin, ellipticine and pentamidine--the so-called "Positive Control" set of ligands) were selected because Surflex and Autodock can successfully reproduce the known crystal structures of these groove binders and intercalators [63]. These compounds were docked to an array of 10 nucleic acids that were constructed *in silico*. The array of nucleic acids are highly diverse and consist of duplex, triplex, and quadruplex DNA and RNA as well as groove and intercalation sites and sequence heterogeneity. Based on the docking results, *in silico* rules were developed to classify the compounds on the basis of their binding mechanism and to assess the sequence and structural selectivity of the compounds. The rules were further tested on several novel triplex and quadruplex binding small molecules (the "Validation" set) that our laboratory has discovered. Finally, the rules were also tested on a 67 set of compounds (the "67 Compound Library" set) for which nucleic acid sequence and structural data for the 10 array of nucleic acid structures was previously acquired by competition dialysis. In summary, we present the development of *in silico* metrics that rationalizes existing data and can predict critical binding information about small molecules *in silico*.

The development of *in silico* rules for predicting small molecule-nucleic acid binding behavior has significant implications for the field of drug discovery. This is a vastly unexplored area of research and there have been almost no efforts to predict the binding mechanisms of small molecules by *in silico* approaches [114]. The development of predictive rules to govern small molecule-nucleic acid interactions will facilitate screening of many small molecules to the *in silico* array of targets to determine binding mode and sequence specificity. This will be a valuable tool to discover new small molecules by virtual screening or alternatively, preempt chemical synthesis of derivatives of known small molecules which is an often expensive and laborious undertaking. Another consideration is that the *in silico* array of 10 nucleic acids is but an initial point for testing. The power of this approach is that the *in silico* screen and library can be expanded or customized as more structures become available *in silico*. In total, we believe this information will allow for the virtual screening of millions of small molecules *in silico* to discover compounds that can bind to a nucleic acid target of interest by a known binding mechanism and sequence specificity. This will provide an essential tool for novel lead compound discovery.

Experimental and Computational Methods

Construction of the In silico Nucleic Acid Library. The first challenge was to build a structurally equivalent *in silico* nucleic acid library compared to the library that was used for competition dialysis for the 67 Compound Set. This array could also be used for the Positive Control and Validation sets as nucleic acids were included in the array that were known to interact with these small molecules. A representative 10 nucleic acid *in silico* library was built and serves as the basis for the docking experiments described here (Table 3). The nucleic acids exhibit a wide variety of structural and sequence diversity. Duplex, triplex and quadruplex nucleic acids are represented as well as an array of binding sites including grooves, intercalation and end-pasting sites (Figure 14).

All of the nucleic acids were either built using Sybyl 8.1 or by direct download from the Protein Data Bank (PDB) database [115]. Unless otherwise noted, all nucleic acids were 12 nucleotides in length. Additionally, the nomenclature used herein for nucleic acids is the following: polydA: polydT consists of one strand Adenine and one strand Thymine while poly(dAdT) consists of alternating Adenine and Thymine nucleotides on each strand. Pure duplex DNAs (poly(dAdT), polydA : polydT, polydG : polydC and poly(dGC)) were built in the B-Form while the RNA-DNA hybrid (polyrA : polydT) and pure RNA (polyrA : polyrU) were built in the A-Form as these are the

Table 3. The following table describes the nucleic acid morphologies, sequences and sites targeted for molecular docking and are identical to those used in competition dialysis. The nomenclature for each nucleic acid will be the nucleic acid identifier in remaining figures and tables in this chapter.

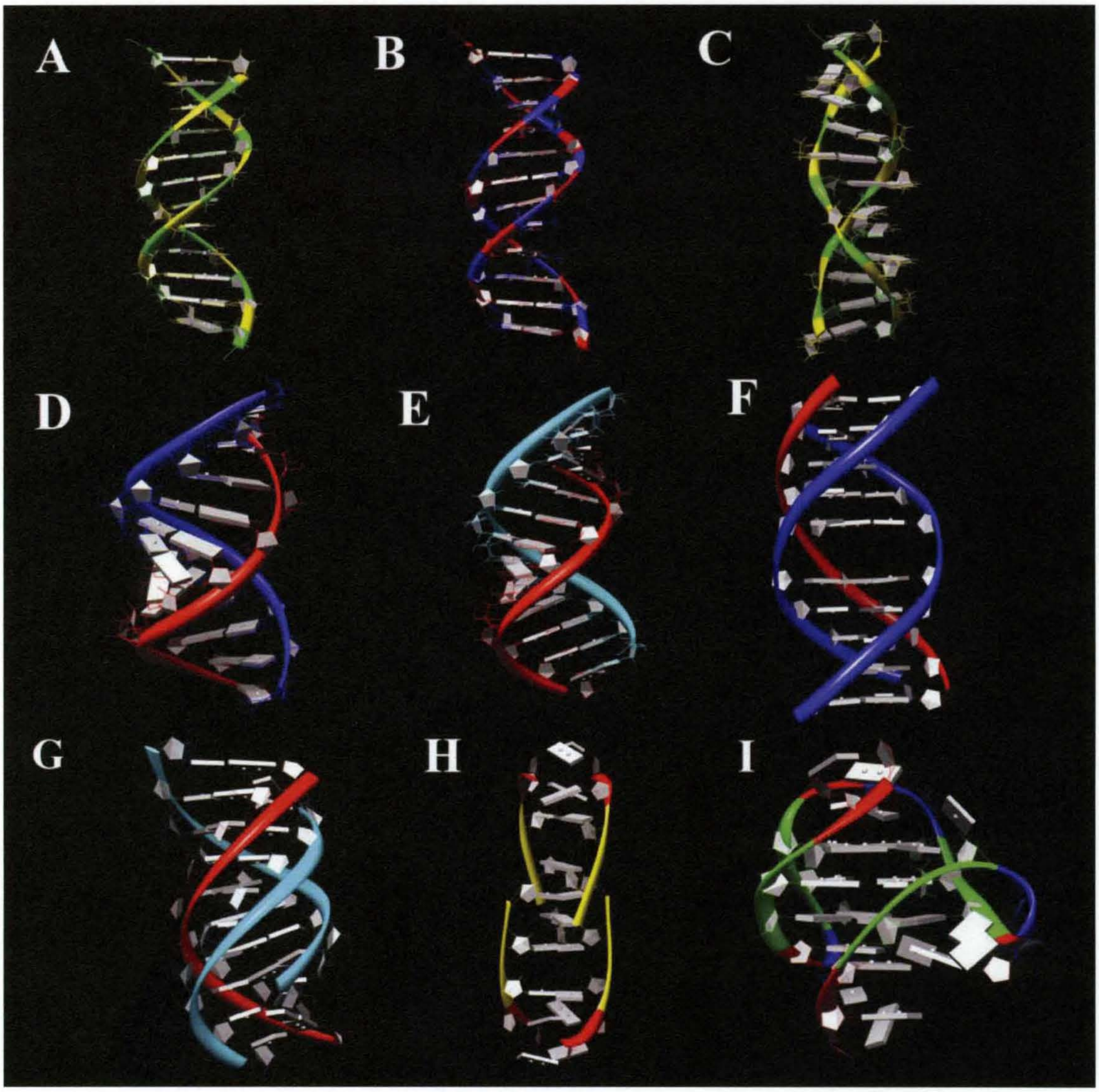
Table 3. The following table describes the nucleic acid morphologies, sequences and sites targeted for molecular docking and are identical to those used in competition dialysis. The nomenclature for each nucleic acid will be the nucleic acid identifier in remaining figures and tables in this chapter.

Nucleic Acid Morphology	**Sequence**	**Targeted Site**	**Nomenclature**
Duplex A form	polyrA : polydT	Major groove	ar2
Duplex A form	polyrA : polyrU	Major groove	au2
Duplex B form	polydA : polydT	Minor groove	ta1
Duplex B form	poly(dAdT)	Minor groove	at2
Duplex B form	polydG : polydC	Minor groove	cg2
Duplex B form	poly(dGC)	Minor groove	gc1
Duplex B form	poly(dGC)	Intercalation	gcit
Duplex Z form[1]	poly(dGC)	Groove 1	zd1
Duplex Z form[1]	poly(dGC)	Groove 2	zd2
Duplex Z form	poly(dGC)	Intercalation	zint
Triplex DNA	poly(dA)-[poly(dT)]$_2$	Minor groove	da1
Triplex RNA	poly(A)-[poly(U)]$_2$	Minor groove	ra2
Triplex DNA	poly(dA)-[poly(dT)]$_2$	Intercalation	dadtdtint
Triplex RNA	poly(A)-[poly(U)]$_2$	Intercalation	raruruint
Imotif[1]	(AACCCC)$_4$	Groove 1	im1
Imotif[1]	(AACCCC)$_4$	Groove 2	im2
Quadruplex	A$_3$GGG(TTAGGG)$_3$A$_2$	Groove 1	1h1
Quadruplex	A$_3$GGG(TTAGGG)$_3$A$_2$	Groove 2	1h2
Quadruplex	A$_3$GGG(TTAGGG)$_3$A$_2$	Groove 3	1h3
Quadruplex	A$_3$GGG(TTAGGG)$_3$A$_2$	Groove 4	1h4
Quadruplex	A$_3$GGG(TTAGGG)$_3$A$_2$	Groove 5	1h5
Quadruplex	A$_3$GGG(TTAGGG)$_3$A$_2$	Intercalation Site 1	1hint1
Quadruplex	A$_3$GGG(TTAGGG)$_3$A$_2$	Intercalation Site 2	1hint2
Quadruplex	A$_3$GGG(TTAGGG)$_3$A$_2$	End Pasting Site 1	1hend1
Quadruplex	A$_3$GGG(TTAGGG)$_3$A$_2$	End Pasting Site 2	1hend2

1: Due to the structural diversity of these DNA, both grooves were targeted

Figure 14. Some of the nucleic acid structures used for the molecular docking experiments: (A) poly(dGC) B Form DNA; (B) poly(dAdT) B Form DNA; (C) poly(dGC) Z Form DNA; (D) polyrA : polydT A Form RNA-DNA hybrid; (E) polyrA : polyrU A Form RNA; (F) poly(dA)-[poly(dT)]$_2$ triplex DNA; (G) poly(A)-[poly(U)]$_2$ triplex RNA; (H) I-Motif and (I) Hybrid-1 Quadruplex DNA. The color scheme is red = Adenine, dark blue = Thymine, green = Guanine, yellow = Cytosine and light blue = Uracil.

Figure 14. Some of the nucleic acid structures used for the molecular docking experiments.

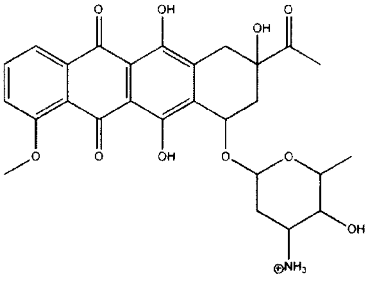


typical forms for these structures *in vivo* [116]. Poly (dGC) was built in the Z-Form and brominated by atom replacement of 50% of the deoxycytosines with 5 – bromodeoxycytosine at alternating positions. The DNA and RNA triplexes, poly(dA)-[poly(dT)]$_2$ and poly(A)-[poly(U)]$_2$, respectively, were constructed using B-type parallel triplex with and without an X-ray structural intercalation site backbone fragments [(PDB) entry 1p20.ent] and minimized holding the heavy atoms fixed. The I-Motif structure of the sequence (AACCCC)$_4$ was downloaded from the PDB [(PDB) entry 1ybl.ent]. The Hybrid-1 quadruplex consisting of the sequence A$_3$GGG(TTAGGG)$_3$A$_2$ was downloaded from the PDB [(PDB) entry 2hy9.ent]. Intercalation sites and end-pasting sites for the Hybrid-1 quadruplex were using methods that will be further described in Chapter V. Briefly, a ligand consisting of four connected purines (a "quaterpurine" ligand) was placed at the site of interest and the nucleic acid was energetically minimized using sequential steepest descent and Polak Ribier Conjugate Gradient Methods iterations, allowing the nucleotides adjacent to the ligand to remain flexible but the remaining structure to be rigid. For all structures, AMBER atom types were assigned using Sybyl 8.1.

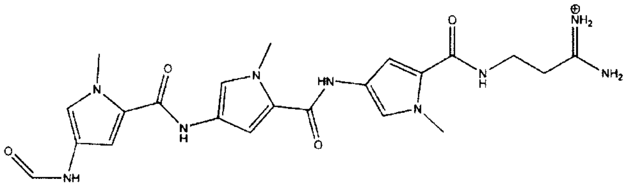
Preparation of the In silico Compound Libraries. Three sets of small molecules were built for the *in silico* molecular docking studies; the "Positive Control" set, the "Validation" set and the "67 Compound Library" set. The Positive Control set consists of the four small molecules which bind to nucleic acids by either groove binding (distamycin and pentamidine) or intercalation (daunorubicin and ellipticine) and were initially used for validation of the molecular docking software Autodock and Surflex for targeting nucleic acid structures (Figure 15) [63, 79-80]. This Positive Control set will be

Figure 15. The Positive Control Set of Ligands used for the molecular docking experiments

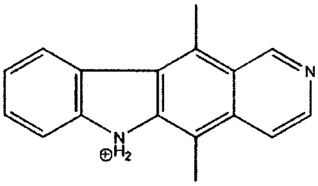
Figure 15. The Positive Control Set of Ligands used for the molecular docking experiments



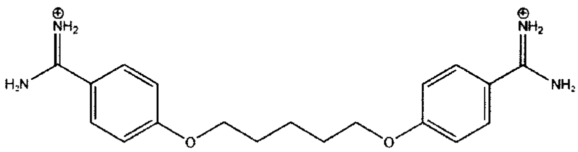
Daunorubicin



Distamycin



Ellipticine



Pentamidine

used to develop the preliminary virtual screening metrics for characterizing the binding mechanism and specificity of small molecules for nucleic acid targets. The second testing set is the Validation set and consists of three novel compounds that we initially discovered *in silico* and demonstrated *in vitro* to bind to either triplex or quadruplex nucleic acid structures. These compounds are referred to as triplex compounds 1 and 2 and the quadruplex compound [64]. The *in silico* rules developed on the Positive Control set were tested on the Validation set to determine if the metrics are predictive of the known binding activity of the compounds in the Validation set. Finally, the third set is the 67 Compound library which were subdivided into 11 smaller sets grouped by chemical similarity (Table 4). These clustered sets possess a range of chemical diversity and contain small molecules that interact with nucleic acids by groove binding, intercalation and end-pasting mechanisms. The 67 Compound set were used for robustness testing of the metrics on an expanded set of small molecules with different binding mechanisms and sequence and structural selectivity. The 67 Compound set was the best data set because we have competition dialysis data on these compounds which will be used as a reference to assess the accuracy of the *in silico* rules for predicting small molecule binding to various sequences and structural nucleic acids. All compounds in these experiments were built using Sybyl 8.1. Charges of the AM1-BCC type were added using the antechamber suite from Amber 8.

Docking of Small Molecule Sets to Nucleic Acid Targets. One of the greatest challenges of molecular docking is that no single program is superior in all facets of a virtual screen. We had previously identified Surflex-Dock 2.4 as superior software for molecular docking and appropriate for large scale virtual screening [63]. To add robustness to our

Table 4. The classification of compounds from the 67 Compound library set grouped by chemical similarity.

Table 4. The classification of compounds from the 67 Compound library set grouped by chemical similarity.

Compound Class Number	**Compound Classification**	**Number of Compounds**
1	Ethidium Bromide Derivatives	9
2	Acridine Derivatives	6
3	Aromatic Diamidine Derivatives	3
4	Cyclic Aromatic Derivatives	6
5	Dibenzophenanthroline Derivatives	5
6	Bis-quinoline Derivatives	8
7	Amidoanthraquinones (aromatic side chains)	9
8	Amidoanthraquinones (non-aromatic side chains)	4
9	Naphthoflavones	2
10	Amidofluorenone Derivatives	4
11	Other Compounds	11

screen; however, we also pursued the development of *in silico* metrics using Autodock 4.0. Autodock is a logical choice for several reasons. First, it is one of the few virtual screening programs with a force field (AMBER) that is parameterized for nucleic acids. We have also previously found Autodock to be comparable to Surflex-Dock at accurately reproducing both groove binding and intercalation crystal structures [63]. Autodock also adds versatility to the virtual screening platform as it is complementary to Surflex with respect to both docking and scoring functions [60]. Finally, Autodock is the most widely cited molecular docking program in the literature, making our findings relevant to the research that is on-going in many laboratories. Importantly, in our previous work, we optimized the docking parameters for these software for docking both minor groove binding and intercalating ligands to nucleic acids. Thus, the same docking parameters used for those experiments were used here. Specifically, for Surflex, the "Multistart 5" option was employed for each ligand and for Autodock, the "Number of Runs" was set to 5 and the "Number of Energy Evaluations" was set to 20,000,000 (2E7). These docking parameters were described in detail in our previous report and were found to reproduce with a high degree of accuracy the crystal structures of a set of small molecule groove binders and intercalators in the Positive control set (Figure 15) [63].

The details of how Surflex and Autodock perform molecular docking have been described in detail in previously [60]. Briefly, Surflex uses protomols that characterize the chemical and spatial properties of the binding site and guides the docking of each small molecule to that site. Protomols were prepared using ligand-based methods against 35 possible binding sites on the 10 nucleic acids, as previously described [63]. The various sites represent groove, intercalation or end-pasting sites that are all possible sites of interaction for each of the docking small molecules. All files were saved in MOL2 format using Sybyl 8.1 prior to molecular docking. Autodock precomputes energy grids around the nucleic acid to characterize the properties of the target [117]. Each ligand is docked and evaluated against the target using a Lamarkian Genetic Algorithm and the top pose was selected as the most energetically stable pose of the ligand with respect to the target. PERL scripts were written to center the Autodock docking energy grids on the center of the Surflex protomol for each site, to best compare the performance of the two programs. Targets were visualized in AutoDockTools to ensure the grid center was centered on the Surflex-Dock protomol. All files were saved in PDBQT format using AutoDockTools. The Autodock grid center and extent of the grid maps for each of the targets can be found in Table 5. For each docking compound, the score for the top ranked pose (the highest docking score) was used for subsequent data analysis. All preparation procedures and docking experiments were performed on a 440 computer IBM server with 2.66GHz Intel(R) Xeon(R) E5430 processors.

Results and Discussion

Initial Docking of the Positive Control Set to the In silico Nucleic Acid Targets.

The initial objective in these studies was to dock the four ligands of the Positive Control set to the array of nucleic acids and determine if the scores would yield insights as to the preferential binding mode (groove binding versus intercalation) and the sequence selectivity of the small molecules. In order to perform docking to these structures, typically a site on the target must be specified to guide the docking. This required generalizing which specific site on each nucleic acid target are “relevant” for small

Table 5. Autodock Grid Properties used for Docking Experiments

Table 5. Autodock Grid Properties used for Docking Experiments

Autodock Target	**Gridcenter (X,Y,Z)**	**Number of Grid Points**
at1	1.463 -0.050 3.149	40 X 40 X 40
at2	-0.541 0.895 4.139	40 X 40 X 40
ta1	4.842 2.005 0.299	40 X 40 X 40
ta2	1.506 -0.159 4.317	40 X 40 X 40
zd1	2.467 -4.137 -1.358	40 X 40 X 40
zd2	5.371 -3.340 -3.319	40 X 40 X 40
cg1	1.100 -3.746 -1.423	40 X 40 X 40
cg2	1.807 1.431 4.152	40 X 40 X 40
gc1	-3.929 -3.230 1.876	40 X 40 X 40
gc2	-2.576 0.570 4.061	40 X 40 X 40
ar1	-1.827 -1.902 0.880	40 X 40 X 40
ar2	-0.412 0.466 8.713	40 X 40 X 40
au1	1.048 -0.070 -0.115	40 X 40 X 40
au2	7.629 2.534 1.076	40 X 40 X 40
da1	-1.574 -1.340 6.844	40 X 40 X 40
da2	9.399 2.736 5.313	40 X 40 X 40
da3	5.451 5.088 0.048	40 X 40 X 40
ra1	6.237 2.560 -1.739	40 X 40 X 40
ra2	5.314 2.801 -0.509	40 X 40 X 40
ra3	-1.385 -1.390 11.110	40 X 40 X 40
im1	5.814 1.498 2.134	40 X 40 X 40
im2	7.521 2.221 0.992	40 X 40 X 40
1h1	5.990 7.152 0.890	40 X 40 X 40
1h2	-2.647 6.861 5.388	40 X 40 X 40
1h3	5.967 0.905 6.972	40 X 40 X 40
1h4	3.016 -0.706 10.441	40 X 40 X 40
1h5	-8.534 -3.914 1.367	40 X 40 X 40
gcit	0.339 -0.334 -2.261	40 X 40 X 40
zint	0.445 0.145 1.473	40 X 40 X 40
dadtdtint	0.214 1.929 -0.129	40 X 40 X 40
raruruint	-0.165 1.302 0.030	40 X 40 X 40
1hint1	0.402 2.698 0.144	40 X 40 X 40
1hint2	1.928 -0.579 -0.217	40 X 40 X 40
1hend1	-1.644 6.950 -0.460	40 X 40 X 40
1hend2	-0.491 5.057 -0.484	40 X 40 X 40

molecule interactions. For example, in the case of distamycin and pentamidine for the Positive Control set, we would expect the highest reported docking scores to be for the minor groove of AT rich B-DNA, as these compounds are known minor groove binders to this sequence [118-119].

Applying this rationale to the other nucleic acid targets, for the RNA and RNA-DNA hybrid structures, the major groove was the target while for triplexes, the minor groove was the initial target. For the quadruplex structure, all grooves were targeted, as the most likely binding sites for compounds with the quadruplex are less clear. Finally, in order to select for molecules in the Positive Control Set that bind by intercalation (daunorubicin and ellipticine), we included multiple intercalation sites in duplex, triplex and quadruplex structures in the nucleic acid library. In total, each of the four compounds in the Positive Control set were docked using both Surflex and Autodock against a total of 25 groove, intercalation and end pasting sites on all 10 nucleic acids (Table 6—highlighted in green). The data were evaluated to determine if the ligands in the Positive Control set could be classified by binding mechanism and sequence specificity based solely on the *in silico* screening results.

The initial docking results for the Positive Control Set are shown in Figures 16 (groove site scores) and 17 (intercalation site scores). In both figures, more positive docking scores for Surflex and Autodock are generally indicative of better binding to the nucleic acid site of interest. Surprisingly, both docking programs appear to dock all of the Positive Control small molecules to all of the sites of interest, regardless of whether the targets are grooves or intercalation sites. Generally, Surflex appears to have higher groove binding scores for distamycin and pentamidine compared to the intercalators

Table 6. The nucleic acids targeted for docking experiments. The targets highlighted in green were used in the original molecular docking experiments while the targets highlighted in yellow were added later to augment the initial experimental design.

Table 6. The nucleic acids targeted for docking experiments. The targets highlighted in green were used in the original molecular docking experiments while the targets highlighted in yellow were added later to augment the initial experimental design.

Nucleic Acid Morphology	Sequence	Targeted Site	Nomenclature
Duplex A form	polyrA : polydT	Minor groove	ar1
Duplex A form	polyrA : polydT	Major groove	ar2
Duplex A form	polyrA : polyrU	Minor groove	au1
Duplex A form	polyrA : polyrU	Major groove	au2
Duplex B form	polydA : polydT	Minor groove	ta1
Duplex B form	polydA : polydT	major groove	ta2
Duplex B form	poly(dAdT)	Major groove	at1
Duplex B form	poly(dAdT)	Minor groove	at2
Duplex B form	polydG : polydC	Major groove	cg1
Duplex B form	poly(dGdC)	Minor groove	cg2
Duplex B form	poly(dGC)	Minor groove	gc1
Duplex B form	poly(dGC)	Major groove	gc2
Duplex B form	poly(dGC)	Intercalation	gcit
Duplex Z form[1]	poly(dGC)	Groove 1	zd1
Duplex Z form[1]	poly(dGC)	Groove 2	zd2
Duplex Z form	poly(dGdC)	Intercalation	zint
Triplex DNA	poly(dA)-[poly(dT)]$_2$	Minor groove	da1
Triplex DNA	poly(dA)-[poly(dT)]$_2$	Minor-Minor groove	da2
Triplex DNA	poly(dA)-[poly(dT)]$_2$	Major groove	da3
Triplex RNA	poly(A)-[poly(U)]$_2$	Major groove	ra1
Triplex RNA	poly(A)-[poly(U)]$_2$	Minor groove	ra2
Triplex RNA	poly(A)-[poly(U)]$_2$	Minor-Minor groove	ra3
Triplex DNA	poly(dA)-[poly(dT)]$_2$	Intercalation	dadtdtint
Triplex RNA	poly(A)-[poly(U)]$_2$	Intercalation	raruruint
Imotif[1]	(AACCCC)$_4$	Groove 1	im1
Imotif[1]	(AACCCC)$_4$	Groove 2	im2
Quadruplex	A$_3$GGG(TTAGGG)$_3$A$_2$	Groove 1	1h1
Quadruplex	A$_3$GGG(TTAGGG)$_3$A$_2$	Groove 2	1h2
Quadruplex	A$_3$GGG(TTAGGG)$_3$A$_2$	Groove 3	1h3
Quadruplex	A$_3$GGG(TTAGGG)$_3$A$_2$	Groove 4	1h4
Quadruplex	A$_3$GGG(TTAGGG)$_3$A$_2$	Groove 5	1h5
Quadruplex	A$_3$GGG(TTAGGG)$_3$A$_2$	Intercalation Site 1	1hint1
Quadruplex	A$_3$GGG(TTAGGG)$_3$A$_2$	Intercalation Site 2	1hint2
Quadruplex	A$_3$GGG(TTAGGG)$_3$A$_2$	End Pasting Site 1	1hend1
Quadruplex	A$_3$GGG(TTAGGG)$_3$A$_2$	End Pasting Site 2	1hend2

1: Due to the structural diversity of these DNA, both grooves were targeted

Figure 16. Surflex-Dock (top) and Autodock (bottom) docking scores for the Positive Control set of compounds, before application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

Figure 16. Surflex-Dock (top) and Autodock (bottom) docking scores for the Positive Control set of compounds, before application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

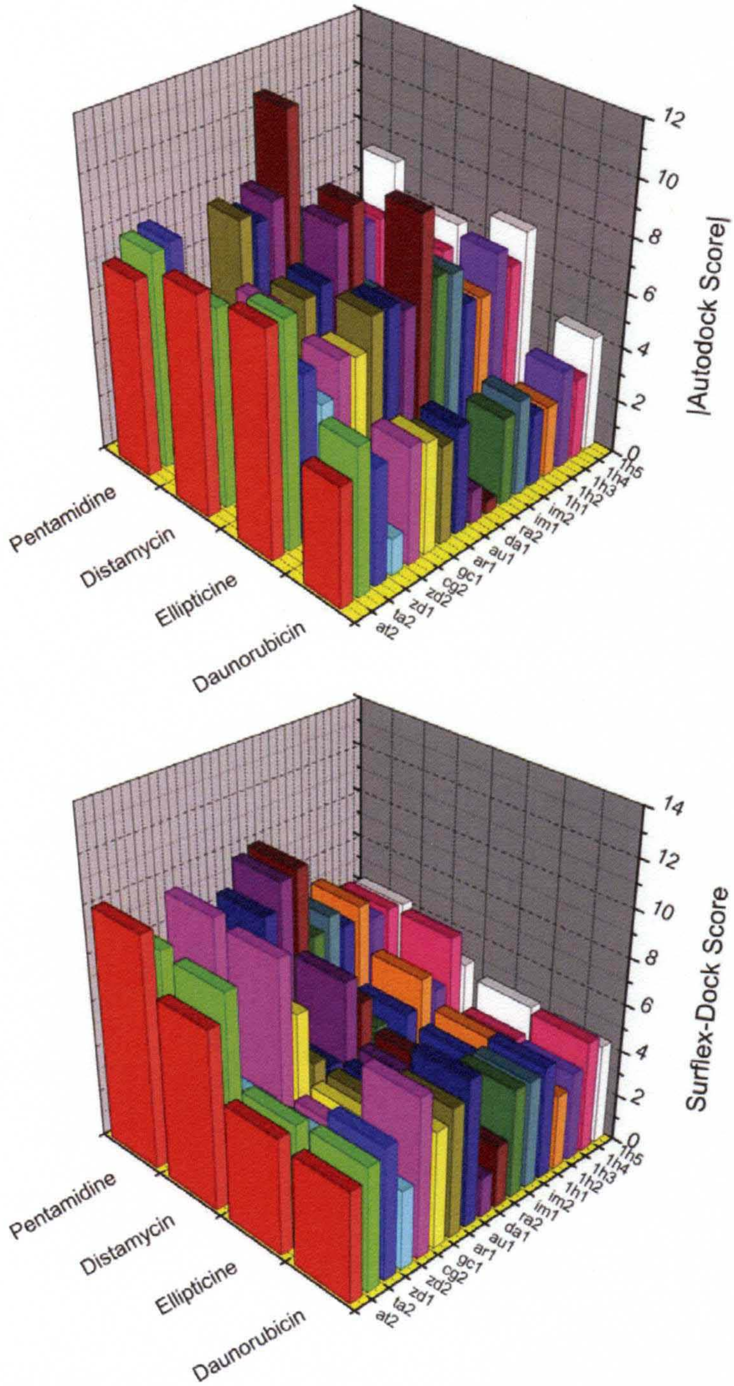
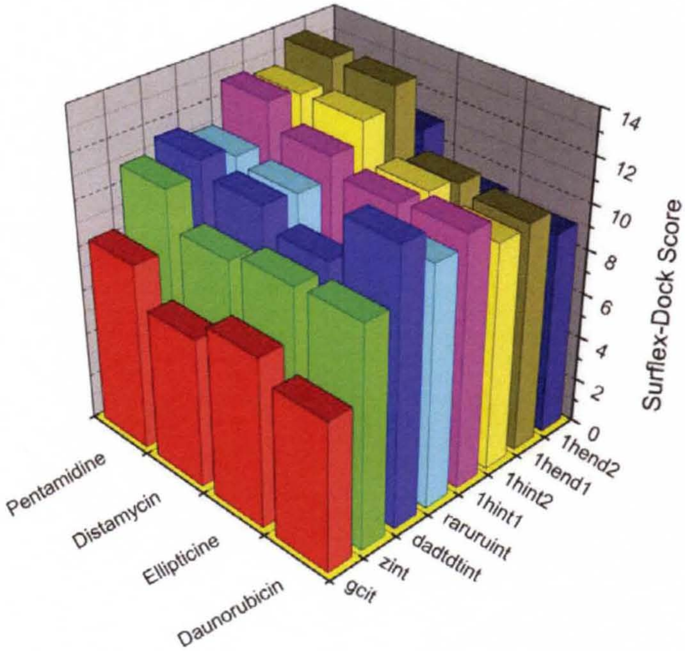
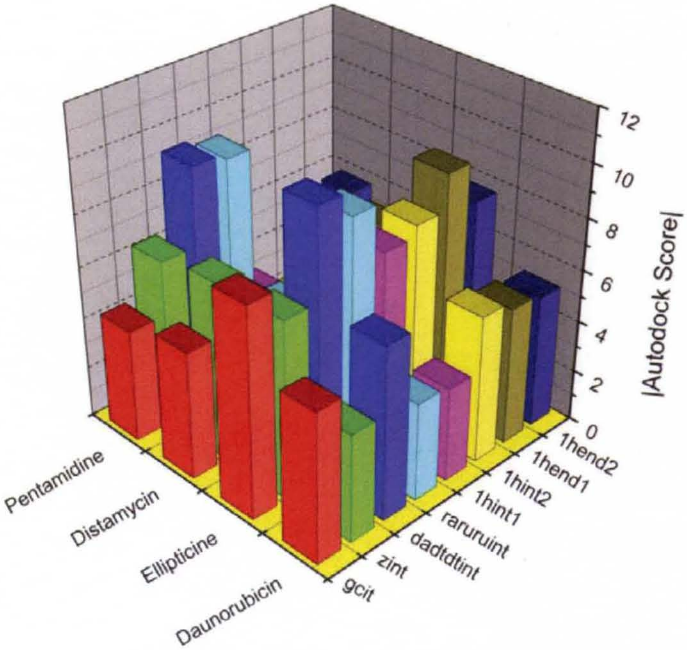


Figure 17. Surflex-Dock (top) and Autodock (bottom) docking scores for the Positive Control set of compounds, before application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.

Figure 17. Surflex-Dock (top) and Autodock (bottom) docking scores for the Positive Control set of compounds, before application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.





daunorubicin and ellipticine, which supports the known preferential groove binding mechanism of these small molecules. However, distamycin and pentamidine also obtain higher Surflex-Dock scores for the intercalation sites than the known intercalators daunorubicin and ellipticine (Figure 17). The trend is less clear with Autodock, and few definitive conclusions can be drawn from these data. Overall, the data as seen in Figures 16 and 17 make it difficult to discern either the "real" binding mechanism or sequence selectivity of the Positive Control Set. This suggests that the initial experimental approach must be augmented and refined to try to elucidate this information from the docking experiments.

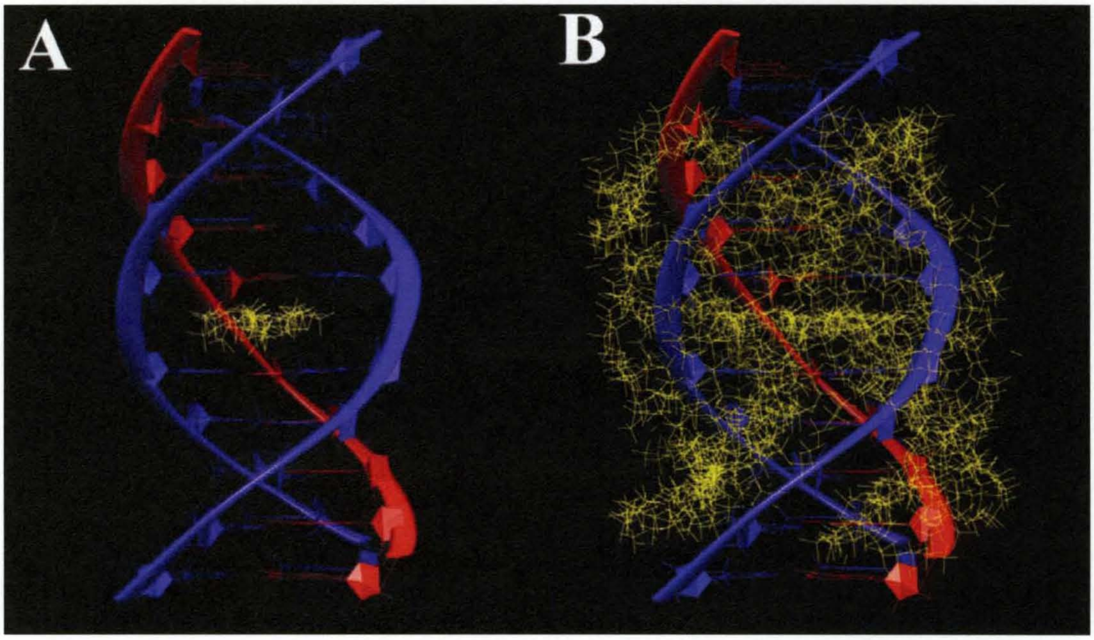
Augmentation of the Initial Positive Control Docking Experiments. The initial docking data for the Positive Control set reveal an important consideration when performing *in silico* docking. The docking software appear to dock compounds to almost any site "successfully" and return some positive value, suggesting that the compound may have some interaction with that site. This suggests that our initial strategy of limiting the initial docking to just the most likely binding site (for example, the minor groove for distamycin) is an over-simplification of what actually occurs *in vitro* in competition dialysis. The positive scores observed at almost all sites suggest that metrics must be developed based on the docking scores to separate the true positive "real" binding data from the "false positives." We briefly revisit how competition dialysis is performed in the hopes of redesigning our strategy to more closely mimic *in silico* what is occurring in competition dialysis *in vitro*.

In competition dialysis, each nucleic acid resides in the retentate of individual dialysis cassettes and is exposed to a ligand that exists in a common dialysate. The ligand has the potential to interact with multiple sites on each nucleic acid in the assay, including both grooves and intercalation sites. The idea of the small molecule having access to multiple sites on a single target led us to hypothesize that multiple sites on a target may non-specifically bind a ligand, so the interaction of a ligand with all of the sites on a target must be taken into account when considering the overall binding of the ligand to the site of interest. Our initial experiments oversimplified this concept as we targeted only the most likely site of interaction of the ligand with the nucleic acid. An example is illustrative. In the case of the triplex DNA, poly(dA)-[poly(dT)]$_2$, we believe it is insufficient to target just an intercalation site to try to identify ligands that act by intercalation (Figure 18A). Instead, all possible binding sites on the triplex must be considered (Figure 18B), including the major grove, minor grove and minor-minor groove. "Non-specific" binding of small molecules to sites other than the one of interest must be accounted for by developing *in silico* metrics to subtract out non-specific interactions. With this in mind, the number of docking sites was expanded from the original 25 (highlighted in green in Table 6) to a total of 35 (including the targets highlighted in yellow in Table 6) to take into account other possible binding sites for ligands on the nucleic acids.

Re-Docking and Metric Development for the Positive Control Set. The Positive Control Set was docked against the expanded library of nucleic acid targets and separate metrics were developed based on the resulting data to classify the compounds in the Positive Control set as either groove binders or intercalators. These groove binding and intercalation metrics will be described in detail below.

Figure 18. (A) The triplex DNA poly(dA)-[poly(dT)]$_2$ with an intercalation site that is designated as the target site by the Surflex-Dock protomol. (B) The same DNA with a protomol covering all available binding sites including the intercalation site, major groove, minor groove and minor-minor groove. Yellow = Surflex protomol. Blue = Thymine. Red = Adenine.

Figure 18. (A) The triplex DNA poly(dA)-[poly(dT)]$_2$ with an intercalation site that is designated as the target site by the Surflex-Dock protomol. (B) The same DNA with a protomol covering all available binding sites including the intercalation site, major groove, minor groove and minor-minor groove. Yellow = Surflex protomol. Blue = Thymine. Red = Adenine.



Groove Binding Metric Development. The metrics developed to determine which compounds bind by groove binding as opposed to intercalation were developed as follows. The metrics seek to take into account "non-specific" binding of a ligand to multiple possible grooves that may exist on a single target. The hypothesis is that intercalators will likely bind with similarly low scores to all grooves while groove binders should bind with much higher scores to their respective minor grooves compared to the other grooves. The difference in scores should allow for discrimination between intercalators and minor groove binders. We have created a Surflex or Autodock "metric score" that takes into account the "non-specific" binding to each nucleic acid target. For example, for duplex B-DNA where we target the minor groove, the raw score for the major groove is subtracted as a "non-specific" interaction, as the minor groove is the typical interaction site for small molecules. The metric for this would therefore be:

$$\mathit{Metric\ score} = \mathit{Score}_{minorgroove} - \mathit{Score}_{majorgroove} - \mathit{CF\ (Correction\ Factor)} \qquad (3)$$

In this example, the compound would be docked against both sites and the metric score would be computed from (3). Note that we still had to include a numerical correction factor (CF) (Surflex: 2.8 and Autodock: 3.0) to differentiate groove binders from intercalators, as we will show later. Using a similar rationale, metrics could be determined for duplex RNA and RNA-DNA hybrids, except in this case, the major groove is where small molecules typically bind, so the smaller minor groove score is subtracted as follows [120]:

$$\mathit{Metric\ score} = \mathit{Score}_{majorgroove} - \mathit{Score}_{minorgroove} - \mathit{CF} \qquad (4)$$

For triplex nucleic acids and quadruplex structures, the situation is more complicated as multiple grooves are present, but a similar rationale applies. In this case, since the minor groove of poly(dA)-[poly(dT)]$_2$ is of interest, the maximum score from either of the other grooves (major or minor-minor) is subtracted as follows:

$$Metric\ score = Score_{minorgroove} - MAX\ (Score_{majorgroove}, Score_{minor\text{-}minorgroove}) - CF \quad (5)$$

The principal idea here is the metric corrects for the docking software's attempt to always find a suitable dock for a small molecule on a target. A complete listing of the groove binding metrics for all sites can be found in Table 7.

Intercalation and End Pasting Metric Development. The metrics developed to discriminate intercalators from groove binders were developed as follows. One lesson learned from our initial docking experiments was that while intercalators had fairly high positive docking scores to *in silico* intercalation sites, unfortunately so did many groove binders. However, we did also find that the true positive groove binders possessed higher groove binding scores to the groove sites than the intercalator ligands. This led us to hypothesize that the true intercalators could be discriminated from the groove binders by subtracting the maximal groove binding score observed across the nucleic acid library from each individual intercalation site score. This would effectively "penalize" groove binders more than intercalators and leave the intercalators with a higher overall net positive score. As we performed with the groove binding metrics, we have created an intercalator metric score. For example, with the triplex DNA poly(dA)-[poly(dT)]$_2$, the metric score for the intercalation site is determined by subtracting the maximal groove site score observed for that compound across all of the grooves (27 groove sites in total) from the intercalation site score. This formula is as follows:

Table 7. The following table describes the groove binding and intercalation metrics that were developed for Surflex and Autodock.

Table 7. The following table describes the groove binding and intercalation metrics that were developed for Surflex and Autodock.

Nucleic Acid Sequence / Nomenclature	**Targeted Site**	**Metric Score (MS) Formulae**[2]
polyrA : polydT / ar2	Major groove	$MS = Score_{major} - Score_{minor} - CF$
polyrA : polyrU / au2	Major groove	$MS = Score_{major} - Score_{minor} - CF$
polydA : polydT / ta1	Minor groove	$MS = Score_{minor} - Score_{major} - CF$
poly(dAdT) / at2	Minor groove	$MS = Score_{minor} - Score_{major} - CF$
polydG : polydC / cg2	Minor groove	$MS = Score_{minor} - Score_{major} - CF$
poly(dGC) / gc1	Minor groove	$MS = Score_{minor} - Score_{major} - CF$
poly(dGC) / gcit	Intercalation	$MS = Score_{intercalation\ site} - MAX\ (Score_{all\ grooves}) - CF$
poly(dGC)[1] / zd1	Groove 1	$MS = Score_{groove1} - Score_{groove2} - CF$
poly(dGC)[1] / zd2	Groove 2	$MS = Score_{groove2} - Score_{groove1} - CF$
poly(dGC) / zint	Intercalation	$MS = Score_{intercalation\ site} - MAX\ (Score_{all\ grooves}) - CF$
poly(dA)-[poly(dT)]$_2$ /da1	Minor groove	$MS = Score_{minor} - MAX(Score_{major}, Score_{minor\text{-}minor}) - CF$
poly(A)-[poly(U)]$_2$ / ra2	Minor groove	$MS = Score_{minor} - MAX(Score_{major}, Score_{minor\text{-}minor}) - CF$
poly(dA)-[poly(dT)]$_2$ / dadtdtint	Intercalation	$MS = Score_{intercalation\ site} - MAX\ (Score_{all\ grooves}) - CF$
poly(A)-[poly(U)]$_2$ / raruruint	Intercalation	$MS = Score_{intercalation\ site} - MAX\ (Score_{all\ grooves}) - CF$
(AACCCC)$_4$[1] / im1	Groove 1	$MS = Score_{groove1} - Score_{groove2} - CF$
(AACCCC)$_4$[1] / im2	Groove 2	$MS = Score_{groove2} - Score_{groove1} - CF$
A$_3$GGG(TTAGGG)$_3$A$_2$ / 1h1	Groove 1	$MS = Score_{groove1} - MAX(Score_{grooves2,3,4,5}) - CF$
A$_3$GGG(TTAGGG)$_3$A$_2$ / 1h2	Groove 2	$MS = Score_{groove2} - MAX(Score_{grooves1,3,4,5}) - CF$
A$_3$GGG(TTAGGG)$_3$A$_2$ / 1h3	Groove 3	$MS = Score_{groove3} - MAX(Score_{grooves1,2,4,5}) - CF$
A$_3$GGG(TTAGGG)$_3$A$_2$ / 1h4	Groove 4	$MS = Score_{groove4} - MAX(Score_{grooves1,2,3,5}) - CF$
A$_3$GGG(TTAGGG)$_3$A$_2$ / 1h5	Groove 5	$MS = Score_{groove5} - MAX(Score_{grooves1,2,3,4}) - CF$
A$_3$GGG(TTAGGG)$_3$A$_2$ / 1hint1	Intercalation Site 1	$MS = Score_{intercalation\ site} - MAX\ (Score_{all\ grooves}) - CF$
A$_3$GGG(TTAGGG)$_3$A$_2$ / 1hint2	Intercalation Site 2	$MS = Score_{intercalation\ site} - MAX\ (Score_{all\ grooves}) - CF$
A$_3$GGG(TTAGGG)$_3$A$_2$ / 1hend1	End Pasting Site 1	$MS = Score_{endpasting\ site} - MAX\ (Score_{all\ grooves}) - CF$
A$_3$GGG(TTAGGG)$_3$A$_2$ / 1hend2	End Pasting Site 2	$MS = Score_{endpasting\ site} - MAX\ (Score_{all\ grooves}) - CF$

1: Due to the structural diversity of these DNA, both grooves were targeted

2: A correction factor (CF) of 2.8 and 3.0 was used for Surflex groove binding and intercalation/end – pasting metrics, respectively. A CF of 3.0 and 0.0 was used for Autodock groove binding and intercalation/end-pasting, metrics, respectively.

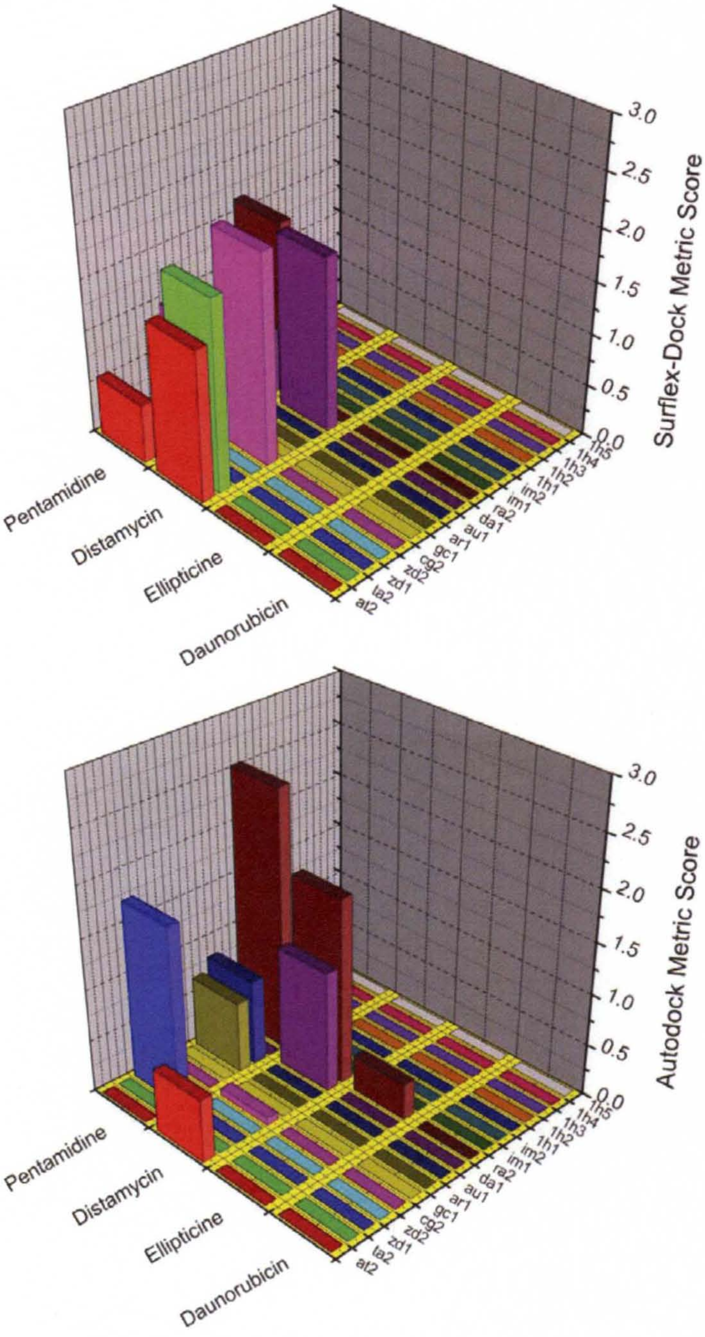
$$Metric\ score = Score_{intercalation\_site} - MAX\ (Score_{all\_nucleicacid\_grooves}) - CF \qquad (6)$$

This formula is applied to all intercalation and end-pasting sites in the nucleic acid library to yield a metric score for these sites. A correction factor of 2.8 was necessary for Surflex to discriminate between groove binders and intercalators, while no correction factor was necessary for Autodock. A detailed list of the metric score formulae can be found in Table 7.

Classification of the Positive Control Library after Metric Application. After completing the re-docking experiments of the Positive Control ligand set to the augmented nucleic acid library containing 35 sites, the groove binder and intercalator metrics were applied to the resulting docking data. The development and application of these *in silico* metrics greatly enhances the trends in the data and makes it possible to classify ligands as groove binders or intercalators, based solely on the transformed *in silico* data. A comparison of the groove binding data prior to metric application (Figure 16) and after application (Figure 19) and intercalation data prior to metric application (Figure 17) and after application (Figure 20) particularly emphasizes this point. It is readily apparent that Surflex has only positive scores for the groove binders pentamidine and distamycin with no scores seen for the intercalators daunorubicin and ellipticine, which is what is expected for the groove sites (Figure 19). A similar general trend is seen with Autodock. While the groove binding metrics can discern groove binders from intercalators, the data show that predicting sequence selectivity is less clear. Surflex appears to do an overall better job compared to Autodock in this area, as Surflex has more positive scores for both the "at" and "ta" sites (AT rich B-DNA) which pentamidine

Figure 19. Surflex-Dock (top) and Autodock (bottom) docking scores for the Positive Control set of compounds, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

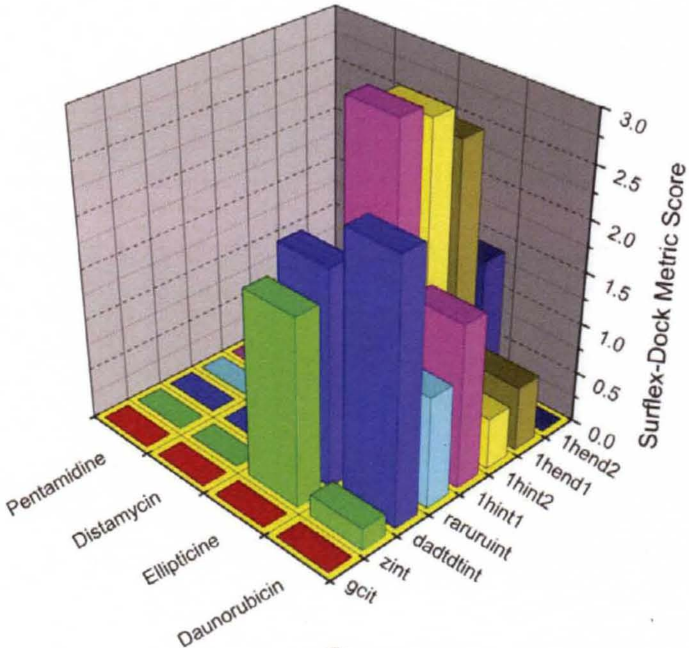
Figure 19. Surflex-Dock (top) and Autodock (bottom) docking scores for the Positive Control set of compounds, <u>after</u> application of the <u>groove binding</u> metrics. The results shown are for the <u>groove binding</u> sites in the nucleic acid library.

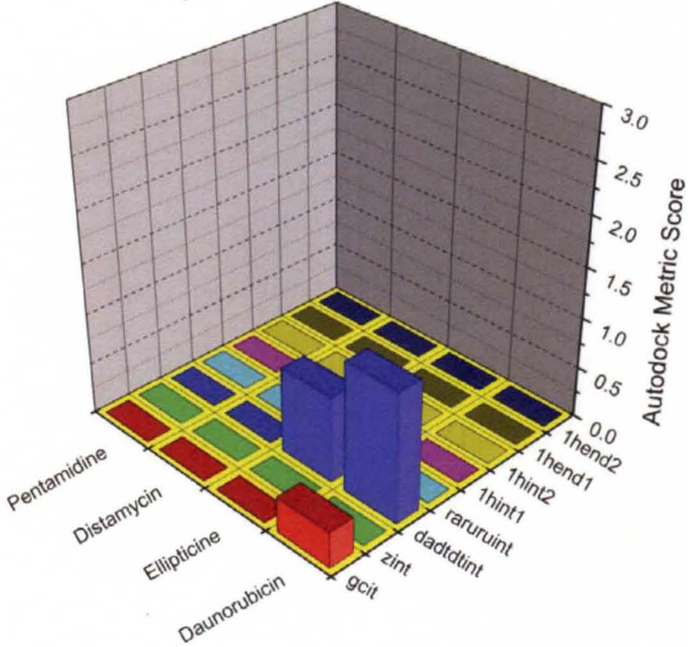


and distamycin are known to bind (Figure 19) [118-119]. It is interesting to observe that the minor groove of triplex DNA and RNA appears to have high scores, even after metric application for both Surflex and Autodock. This suggests that more subtle differences in structure are perhaps difficult to discriminate with the software and groove binding metrics that have been developed here. With the success of the groove binding metrics at preferably identifying groove binders over intercalators, the next question is whether the intercalator metrics could select out the intercalating ligands, daunorubicin and ellipticine. The intercalator metrics were developed and applied to the Positive Control set and the results are shown in Figure 20. For both Surflex and Autodock, after application of these metrics, only positive scores are seen for the intercalation sites with the intercalators daunorubicin and ellipticine, while no scores are seen for the groove binding ligands, distamycin and pentamidine. Surflex appears to have more overall positive scores for different types of intercalation sites, suggesting that prediction of *in silico* sequence selectivity may be more problematic. Surprisingly, for Surflex, neither daunorubicin nor ellipticine were predicted to bind to the "gcit" duplex intercalation site after application of the metrics. This is significant as the "gcit" target represents a "typical" intercalation site consisting of duplex B-DNA with a GC flanking sequence. We believe that this could be in part due to the way Surflex scores docked poses which is predominately shape and contact based. With triplex and quadruplex intercalation sites, there is more surface area present which could artificially elevate the Surflex score and unfairly penalize smaller intercalation sites such as "gcit." Interestingly, Autodock appears to predict intercalation of the intercalator Daunorubicin to "gcit." (Figure 20). The fact that Autodock does appear more sequence selective may be because Autodock

Figure 20. Surflex-Dock (top) and Autodock (bottom) docking scores for the Positive Control set of compounds, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.

Figure 20. Surflex-Dock (top) and Autodock (bottom) docking scores for the Positive Control set of compounds, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.





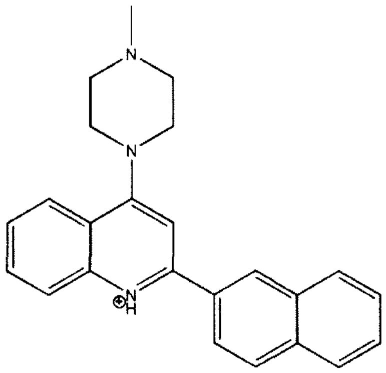
operates using a semi-empirical Amber Force Field which has been appropriately parameterized for DNA and thus may be more appropriate for our uses here.

In summary, the newly developed groove binding and intercalation metrics appear capable of generally predicting the binding mechanism of the Positive Control set of ligands. The question remains whether the metrics that were developed here would also be predictive for the binding mechanism of other small molecules that we have discovered as well as when larger and more diverse chemical compound sets are tested, such as the 67 compound set. Additionally, the question of whether the metrics can predict sequence selectivity will be further evaluated by looking at the 67 Compound Library.

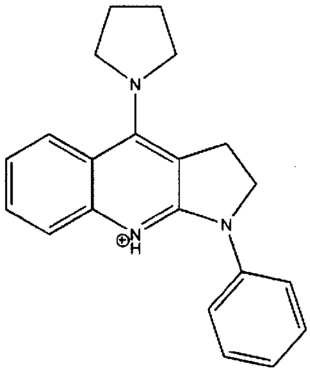
Application of Metrics to the Validation Set. The Validation Set consists of two triplex DNA and one G-quadruplex DNA binding compound that our lab discovered using *in silico* based methods, as we will describe in detail in the next two chapters (Figure 21) [64]. Using the same metrics developed on the Positive Control Set, we sought to determine how the metrics would classify the mechanism of binding (groove binding versus intercalation) of these compounds as well as the predicted sequence selectivity of the compounds. This test of the metrics was valuable as we have already biophysically characterized the binding behavior of these compounds and can compare the predicted *in silico* behavior with the known binding behavior *in vitro*. The triplex compounds were found to intercalate selectively into the triplex poly(dA)-[poly(dT)]$_2$ by Induced Circular Dichroism (ICD) experiments [64]. As we will show in Chapter V, the quadruplex compound binds to the human telomeric quadruplex AG$_3$(TTAGGG)$_3$ by end pasting and

Figure 21. The Validation Set of Ligands used for the molecular docking experiments

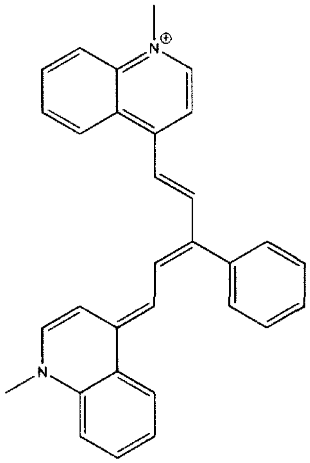
Figure 21. The Validation Set of Ligands used for the molecular docking experiments



Triplex Compound 1



Triplex Compound 2



Quadruplex Compound

possibly intercalation as shown by ICD and a Thiazole Orange Fluorescent Intercalation Displacement Assay (TO-FID).

The Surflex and Autodock results before application of the groove binder and intercalator metrics can be seen in Figures 22 and 23, respectively. As with the Positive Control ligands, the raw data make it difficult to determine whether the triplex and quadruplex compounds are groove binders or intercalators. After applying the groove binding metrics, there are very few positive scores observed for any of the groove interaction sites (Figure 24), particularly for Surflex, which is expected, as the compounds are known intercalators. The Autodock data is somewhat less clear, but generally there are few groove binding scores overall, suggesting that groove binding is not the primary mechanism of action (Figure 24). After application of the intercalator metrics, the Surflex data show prominent positive scores in multiple intercalation sites for the triplex and quadruplex compounds (Figure 25). This suggests that intercalation is the primary mechanism of action of these compounds, which is also consistent with the known binding properties of these compounds. Autodock also shows positive scores particularly in the triplex intercalation sites for the triplex and quadruplex compounds except for triplex compound 1, for which no scores are present. Additionally, Autodock predicts that the quadruplex binding ligand will preferably intercalate into the triplex structure, but this binding behavior has not been biophysically determined. Overall, however, the application of the metrics to the Validation set suggests that the metrics (particularly with Surflex) are able to generally classify known intercalators correctly, but it remains a challenge to also predict sequence selectivity of these ligands. The next

Figure 22. Surflex-Dock (top) and Autodock (bottom) docking scores for the Validation set of compounds, <u>before</u> application of the <u>groove binding</u> metrics. The results shown are for the <u>groove binding</u> sites in the nucleic acid library.

Figure 22. Surflex-Dock (top) and Autodock (bottom) docking scores for the Validation set of compounds, before application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

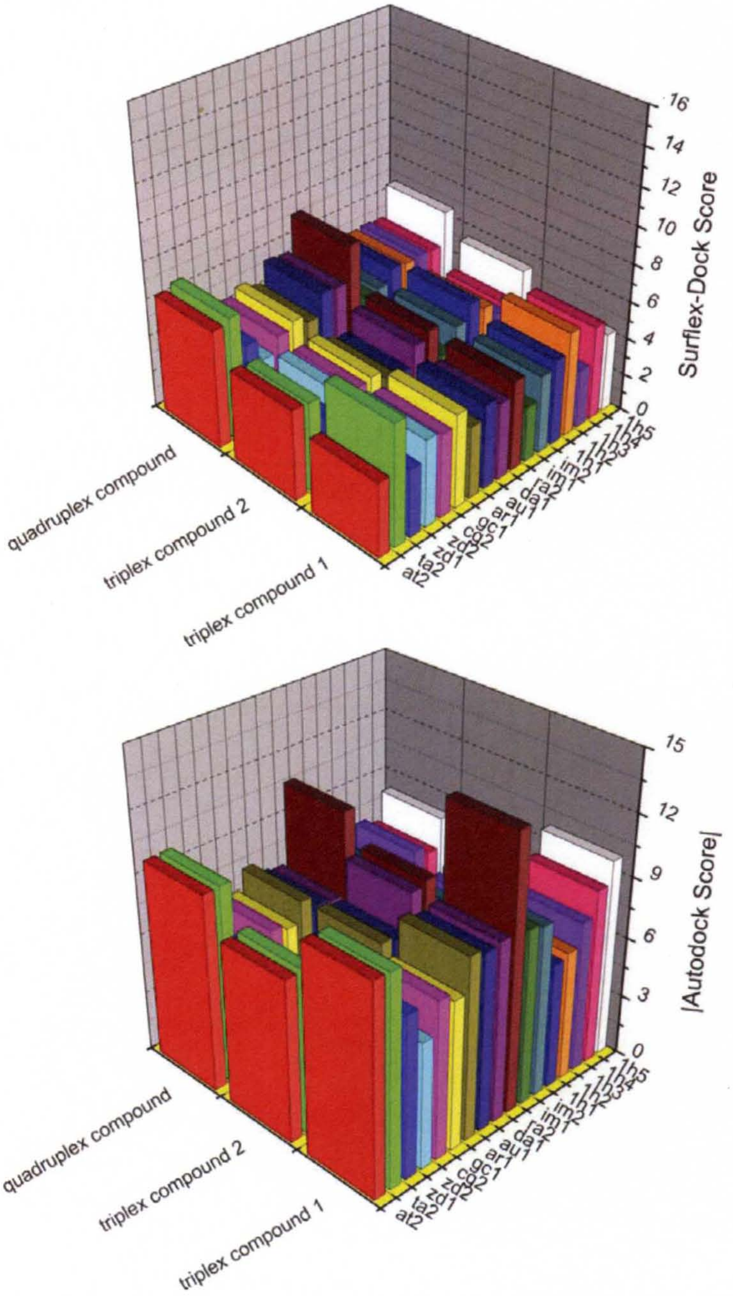


Figure 23. Surflex-Dock (top) and Autodock (bottom) docking scores for the Validation set of compounds, before application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.

Figure 23. Surflex-Dock (top) and Autodock (bottom) docking scores for the Validation set of compounds, before application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.

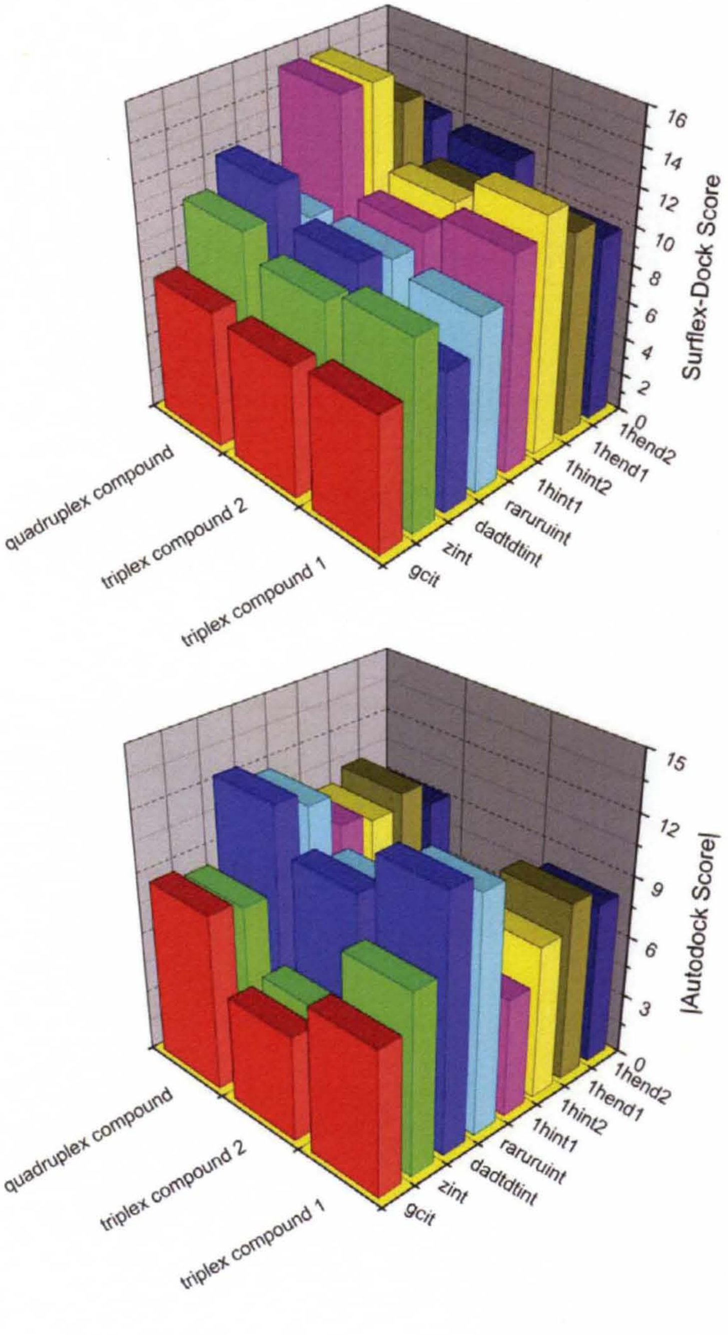


Figure 24. Surflex-Dock (top) and Autodock (bottom) docking scores for the Validation set of compounds, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

Figure 24. Surflex-Dock (top) and Autodock (bottom) docking scores for the Validation set of compounds, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

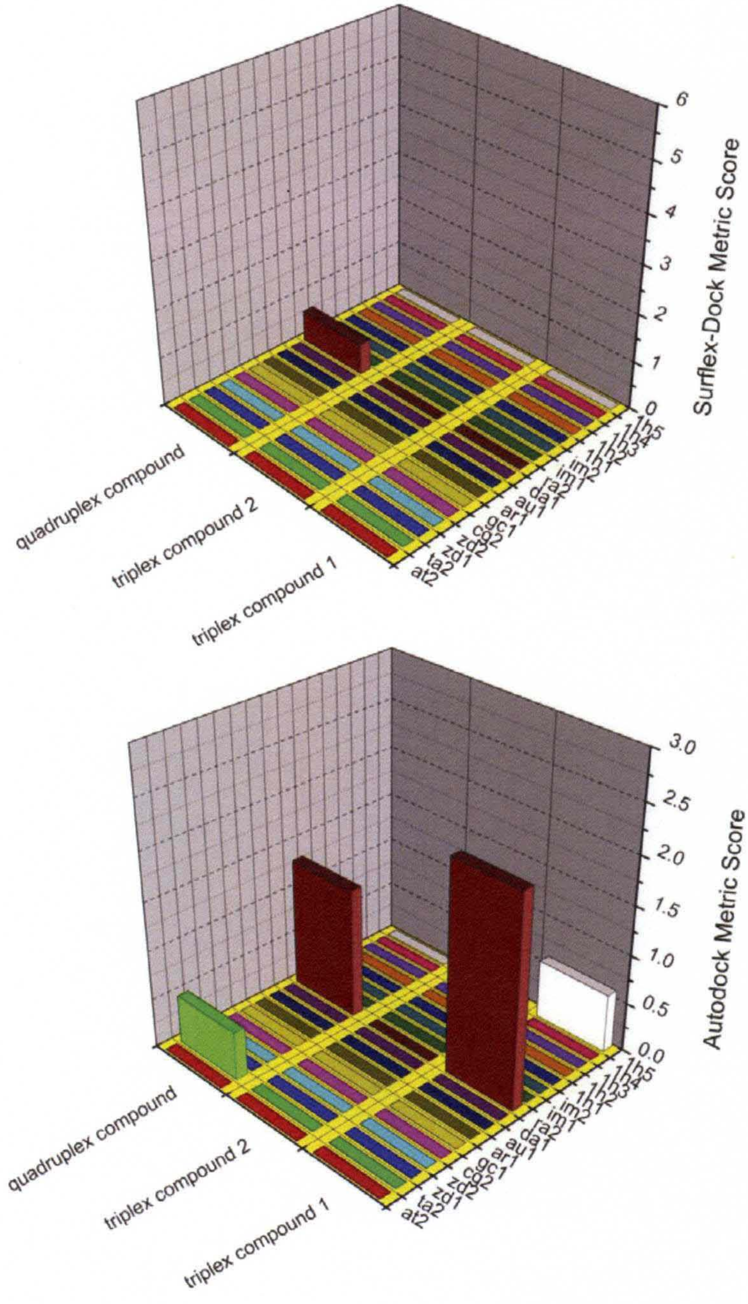
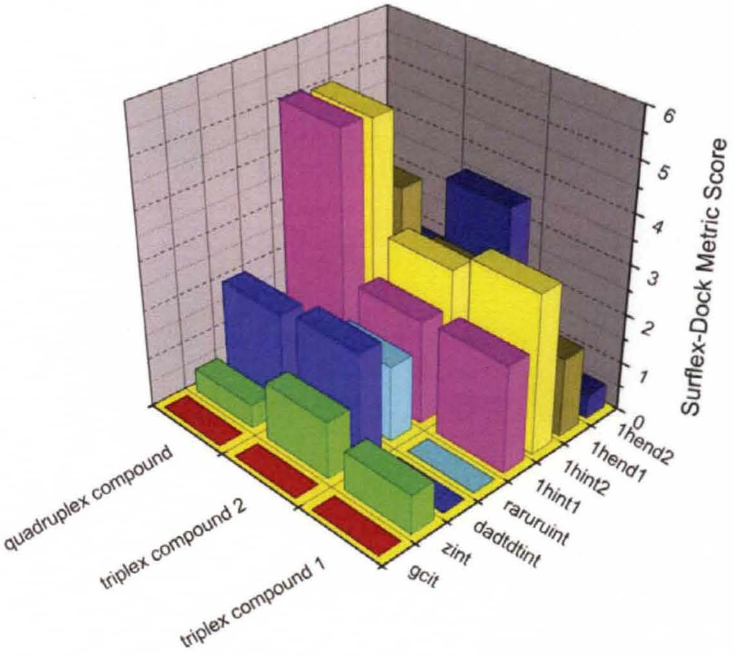
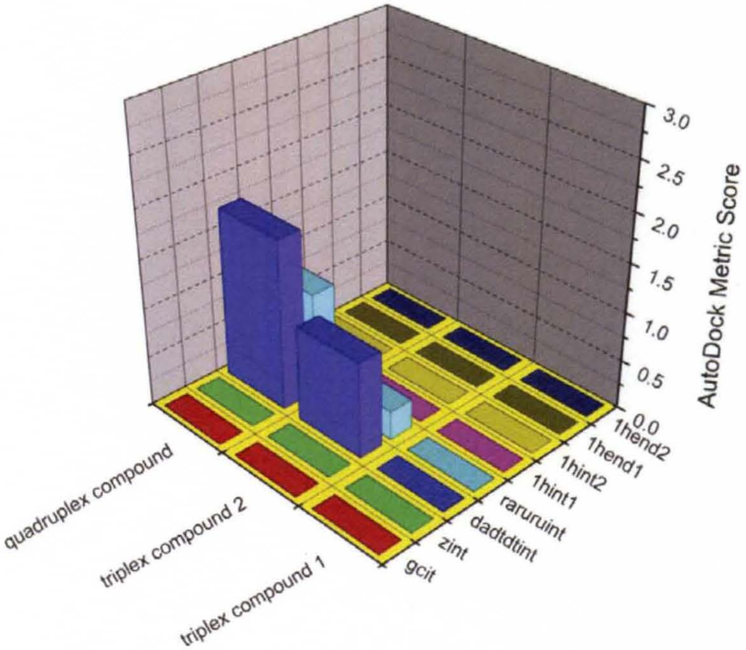


Figure 25. Surflex-Dock (top) and Autodock (bottom) docking scores for the Validation set of compounds, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.

Figure 25. Surflex-Dock (top) and Autodock (bottom) docking scores for the Validation set of compounds, <u>after</u> application of the <u>intercalation</u> metrics. The results shown are for the <u>intercalation</u> sites in the nucleic acid library.





question is whether these metrics can be applied to the 67 Compound set to correctly classify the mechanism of action of these ligands. Moreover, the 67 compound set is particularly valuable here to assess the sequence selectivity, as competition dialysis data exists for each compound against the array of nucleic acids that was used for the *in silico* studies.

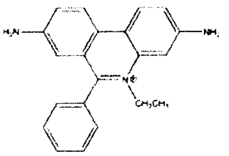
Application of Metrics to the 67 Compound Set. The 67 Compound Set of ligands consists of both groove binders and intercalators with unique nucleic acid sequence selectivity determined by competition dialysis. The ligands have varying length, aromaticity and chemical features that make this diverse set of compounds appropriate for testing the metrics that have been developed on the Positive Control set and tested on the Validation set. For ease of comparison, the compounds have been grouped into sets of chemically similar compounds as shown in Table 4. The structures of the compounds can be found in Figure 26. For each class of compounds, the known binding mechanism and sequence selectivity as reported in the literature and determined by competition dialysis will be briefly discussed below. The metrics that were developed will then be applied to each class of compounds to determine if the compounds act as groove binders or intercalators. Finally, the molecular docking data after application of the metrics will then be compared to the known sequence selectivity data determined by competition dialysis to assess the accuracy of the metrics for predicting sequence selectivity.

Ethidium Bromide Derivatives. Ethidium Bromide is the quintessential DNA "classical" intercalating small molecule that binds non-specifically to many types of DNA and RNA [10, 112, 121-124]. It possesses the "typical" structure of the known nucleic acid intercalators; a flat, planar aromatic surface that can facilitate stacking between adjacent

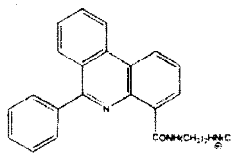
Figure 26. The Subsets of the 67 Compound Set of Ligands used for the molecular docking experiments.

Figure 26. The Subsets of the 67 Compound Set of Ligands used for the molecular docking experiments.

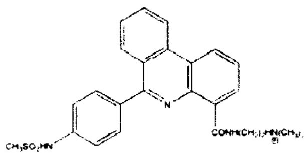
1) Ethidium Bromide Derivatives



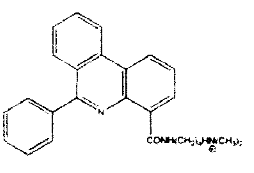
ethidium bromide



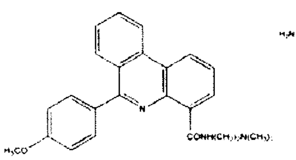
23684



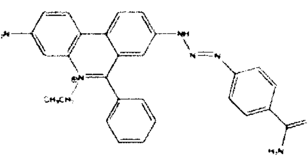
23717



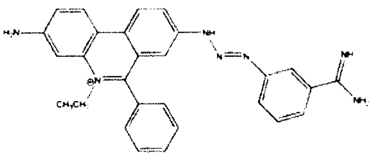
23797



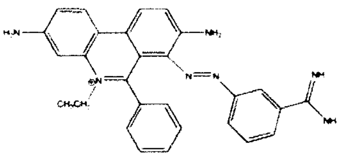
23847



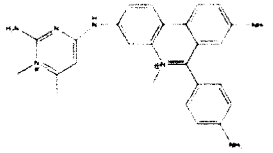
7768



8361

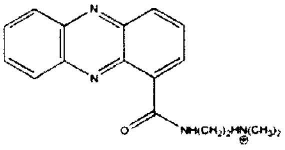


8362

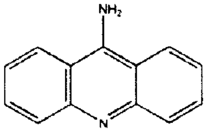


9944

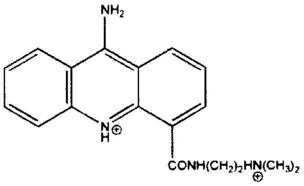
2) Acridine Derivatives



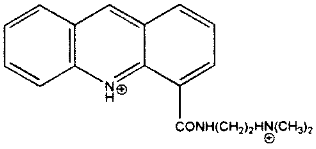
23393



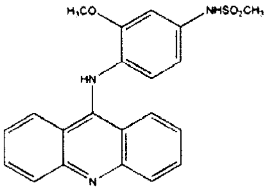
9aminoacridine



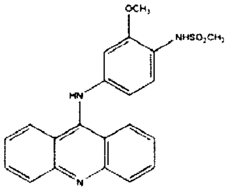
aac



ac

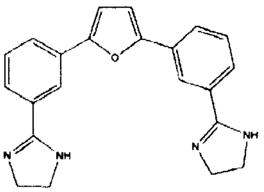


m-amsa

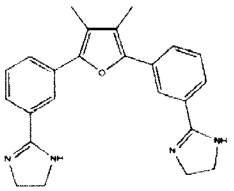


o-amsa

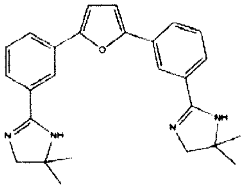
3) Aromatic Diamidine Derivatives



Db361

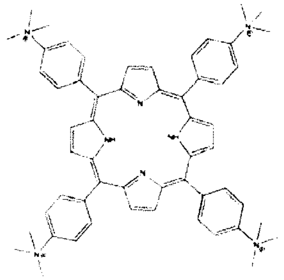


Db443

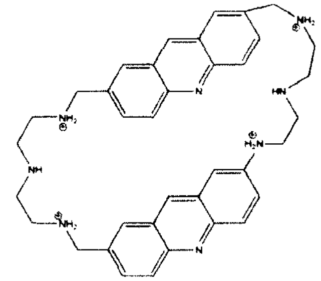


Db471

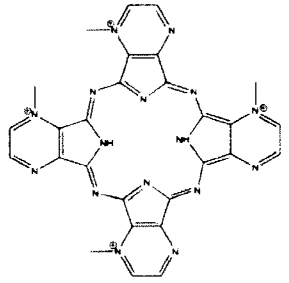
4) Cyclic Aromatic Derivatives



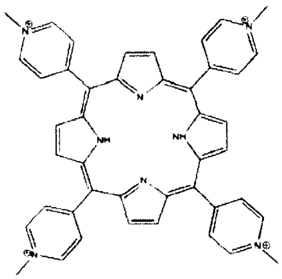
tetrakisporphine



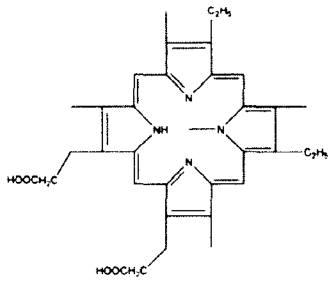
bisa



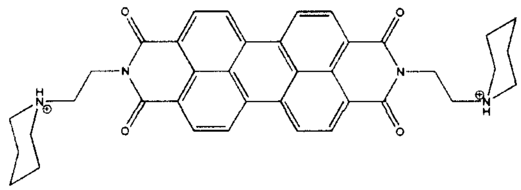
Hoa



mesotetrakisporphine

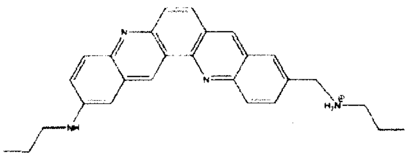


Nmm580

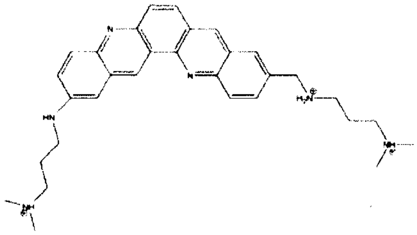


PIPER

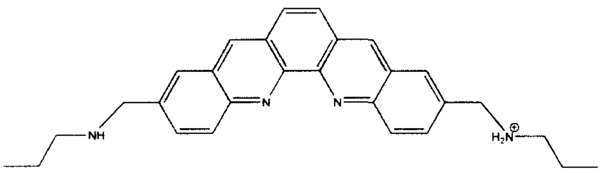
5) Dibenzophenanthroline Derivatives



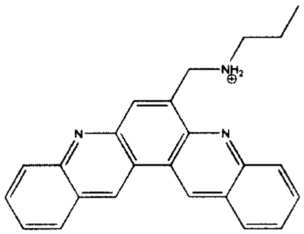
Mmq1



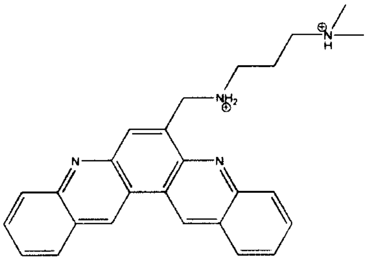
Mmq3



Moq1

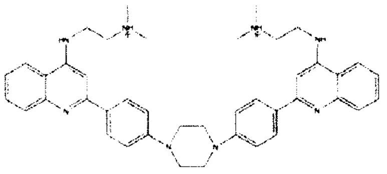


Mpq2

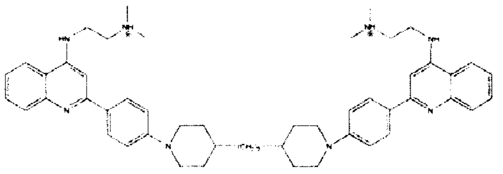


Mpq3

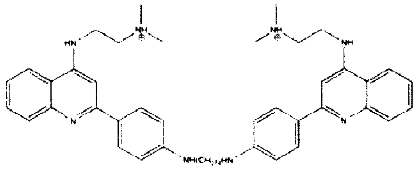
6) Bis-quinoline Derivatives



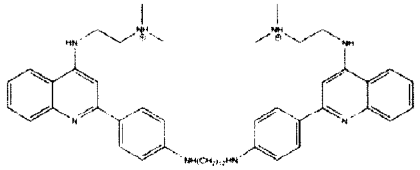
Ms105



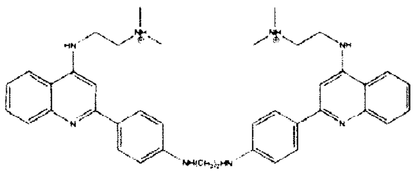
Ms161



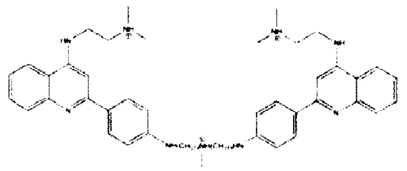
Ms166



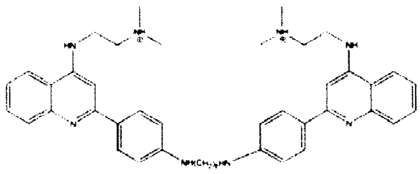
Ms167



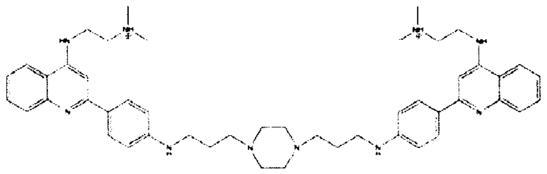
Ms168



Ms170

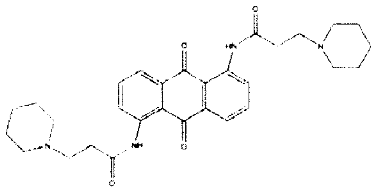


Ms171

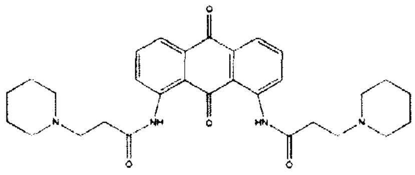


Ms172

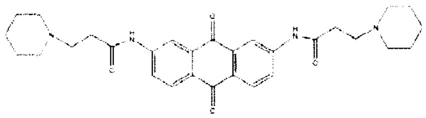
7) Amidoanthraquinones (aromatic side chains)



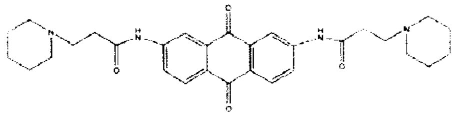
Pjp10



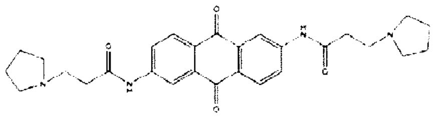
Pjp42



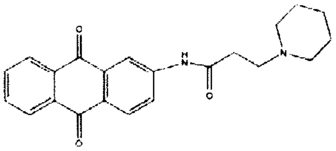
Pjp57



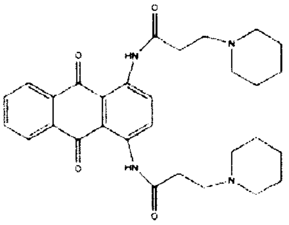
Pjp66



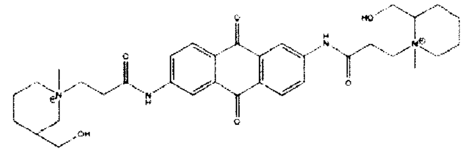
Tcj78



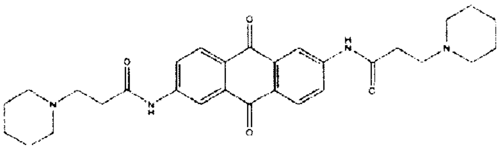
Tcj2107



Tcj74

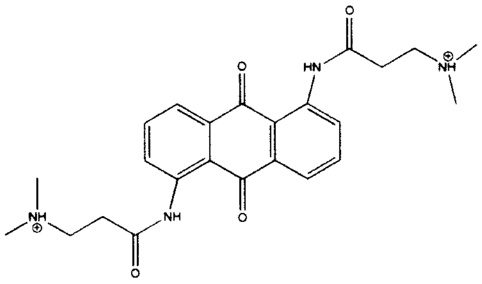


Telominh1

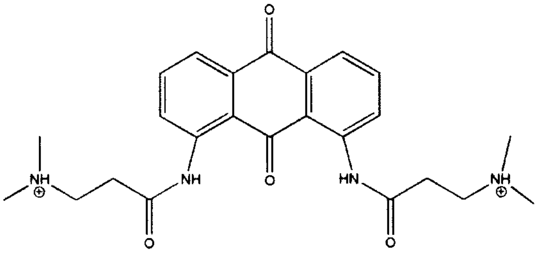


Telominh5

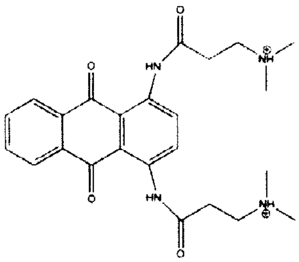
8) Amidoanthraquinones (non-aromatic side chains)



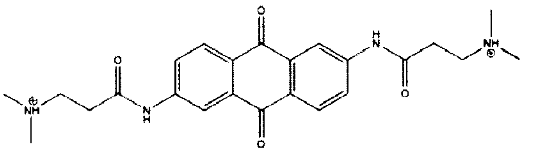
Pjp19



Pjp51

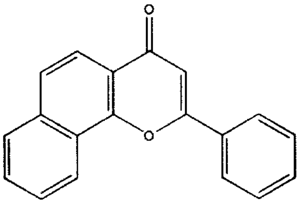


Tcj62

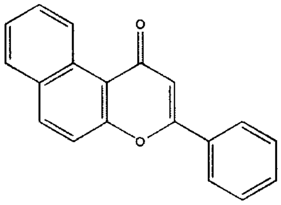


Tcj69

9) Naphthoflavones

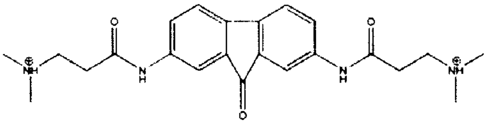


Alpha Naphthoflavone

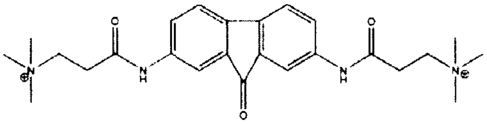


Beta Naphthoflavone

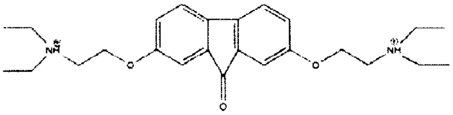
10) Amidofluorenone Derivatives



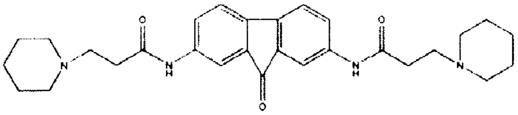
Pjp114



Pjp116

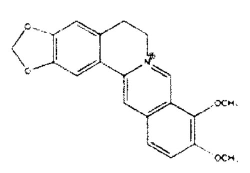


Pjp71

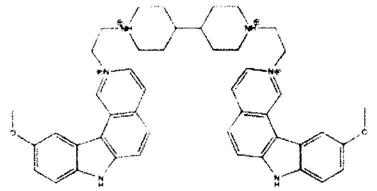


Pjp74

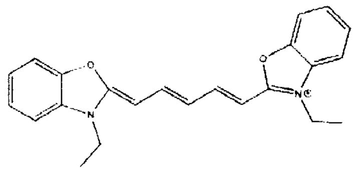
11) Other Compounds



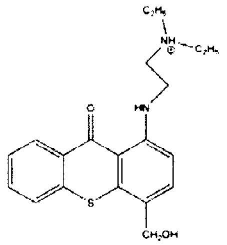
Berberine



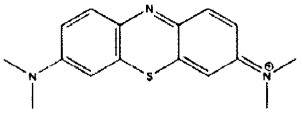
Ditercalinium



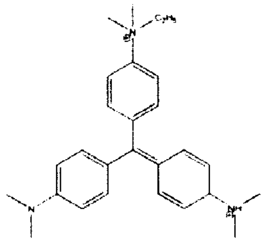
DODC



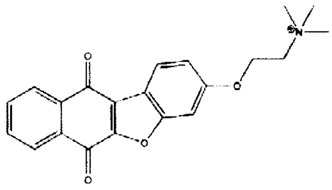
Hycanthone



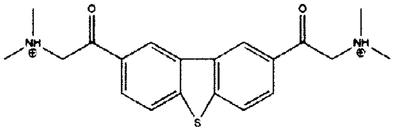
Methylene Blue



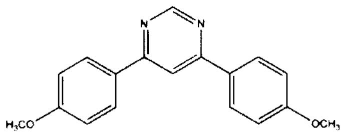
Methylgreen



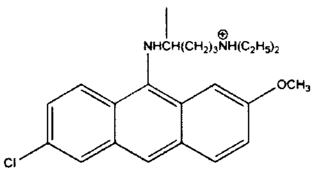
Pjp407



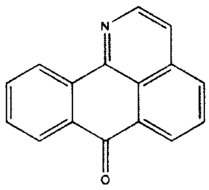
Pjp72



Pm008



Quinacarine



Sampangine

base pairs. The ethidium bromide derivatives (Figure 26) may act by both intercalation and groove binding, as the primary aromatic system may intercalate and the substituents may subsequently interact with the grooves, but generally they act primarily by intercalation [122]. The competition dialysis data (Figure 27) demonstrates this type of promiscuous binding of ethidium bromide to almost all of the structures in the competition dialysis assay.

The Surflex and Autodock *in silico* data, after application of the metrics, show very few positive scores for groove sites (Figure 28) while many positive scores for the intercalation sites (Figure 29). This suggests that the *in silico* screen classifies the ethidium bromide derivatives as predominantly intercalators, which is their known mechanism of action. The *in silico* results for sequence specificity are more variable. Surflex generally has higher scores for the compounds binding to the various quadruplex intercalation and end-pasting sites while Autodock has higher scores for the triplex DNA intercalation sites (Figure 29). This data is consistent with the competition dialysis data that shows many of these compounds binding to both of the triplex and quadruplex DNA forms (Figure 27). However, there is almost a complete absence of predicted binding to duplex DNA, which is in contrast to the binding data from competition dialysis. Overall, the metrics can generally successfully classify the ethidium bromide derivatives as intercalators and but it is generally more challenging to predict sequence preference.

Acridine Derivatives. The acridine derivatives (Figure 26) are another group of classical DNA intercalating agents [124]. Compounds in this chemical family are of interest because of their potent anti-bacterial and anti-neoplastic activity [125-126]. The potent intercalation activity of the acridines has contributed to the development of "hybrid"

Figure 27. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the ethidium bromide derivative set from the 67 Compound Set [10].

Figure 27. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the ethidium bromide derivative set from the 67 Compound Set [10].

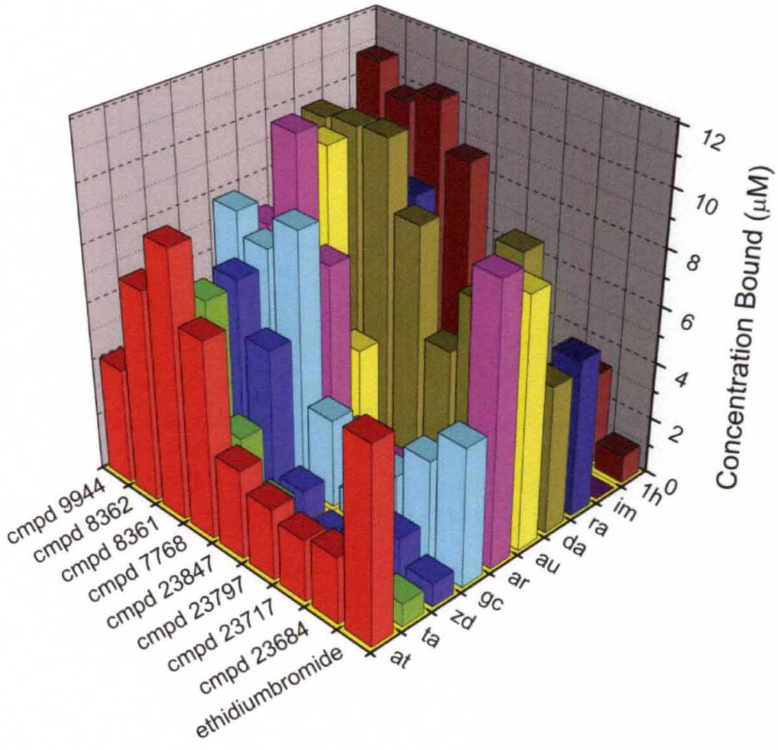


Figure 28. Surflex-Dock (top) and Autodock (bottom) docking scores for the ethidium bromide derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

Figure 28. Surflex-Dock (top) and Autodock (bottom) docking scores for the ethidium bromide derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

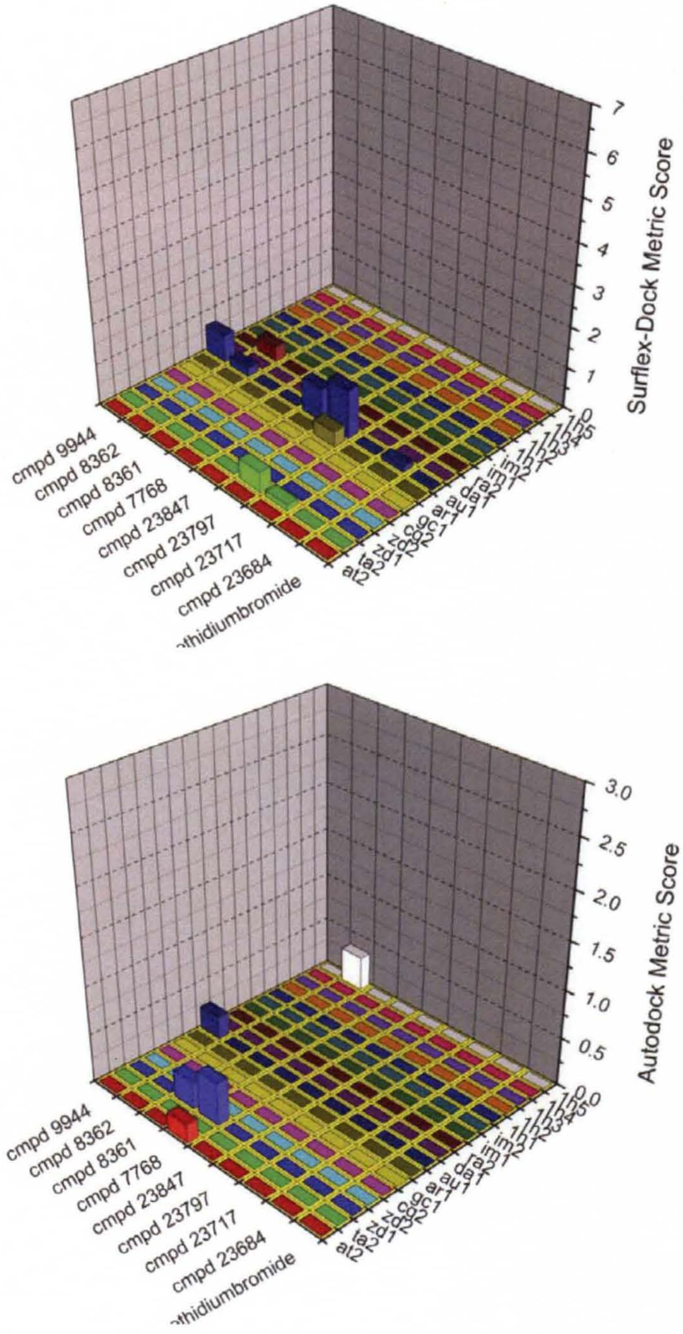
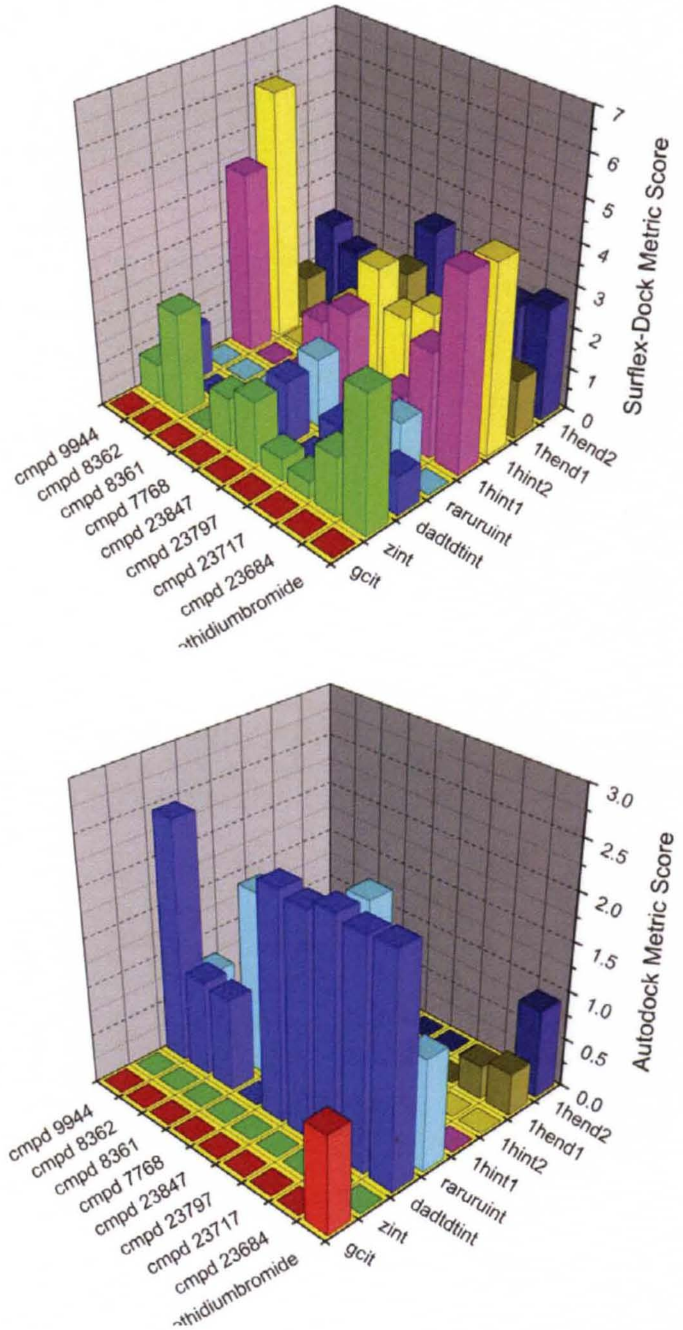


Figure 29. Surflex-Dock (top) and Autodock (bottom) docking scores for the ethidium bromide derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.

Figure 29. Surflex-Dock (top) and Autodock (bottom) docking scores for the ethidium bromide derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.



molecules or "threading intercalators", by attaching DNA minor groove binding agents (such as netropsin) to a molecule with an acridine core scaffold (such as amsacrine) to impart both intercalation and groove binding properties to a single molecule [106]. The goal of this approach is to create a high affinity ligand with sequence specificity and these efforts have been modestly successful [106]. Another member of this class is amsacrine (also known as m-amsa) which has been shown to bind to topoisomerase II [124, 127-128]. Another example is BRACO-19 (a 3,6,9 trisubstituted acridine), one of the most potent G-quadruplex binding ligands discovered to date, consists of an aromatic acridine scaffold that is thought to end-stack with the G-quadruplex, and three "arms" that may bind the grooves and provide quadruplex specificity [126, 129].

Application of the groove binding (Figure 30) and intercalation (Figure 31) metrics shows most of the positive *in silico* scores present in the intercalation sites, supporting the known intercalation of the acridines. Interestingly, Autodock also predicts some triplex groove binding for the aac and ac compounds, which is possible given that these two compounds possess an aromatic core that can intercalate as well as an extended substituent that may also occupy available grooves of the nucleic acid. The competition dialysis data shows the acridines binding predominately to AT and GC rich duplex and triplex DNA structures and sequences (Figure 32). Surflex predicts intercalation into duplex, triplex and quadruplex nucleic acids while Autodock predicts intercalation mostly into triplex DNA and some quadruplex DNA (Figure 31). <u>For this class of compounds, the metrics generally successfully predict the mechanism of action with moderate success in predicting sequence selectivity.</u>

Figure 30. Surflex-Dock (top) and Autodock (bottom) docking scores for the acridine derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

Figure 30. Surflex-Dock (top) and Autodock (bottom) docking scores for the acridine derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

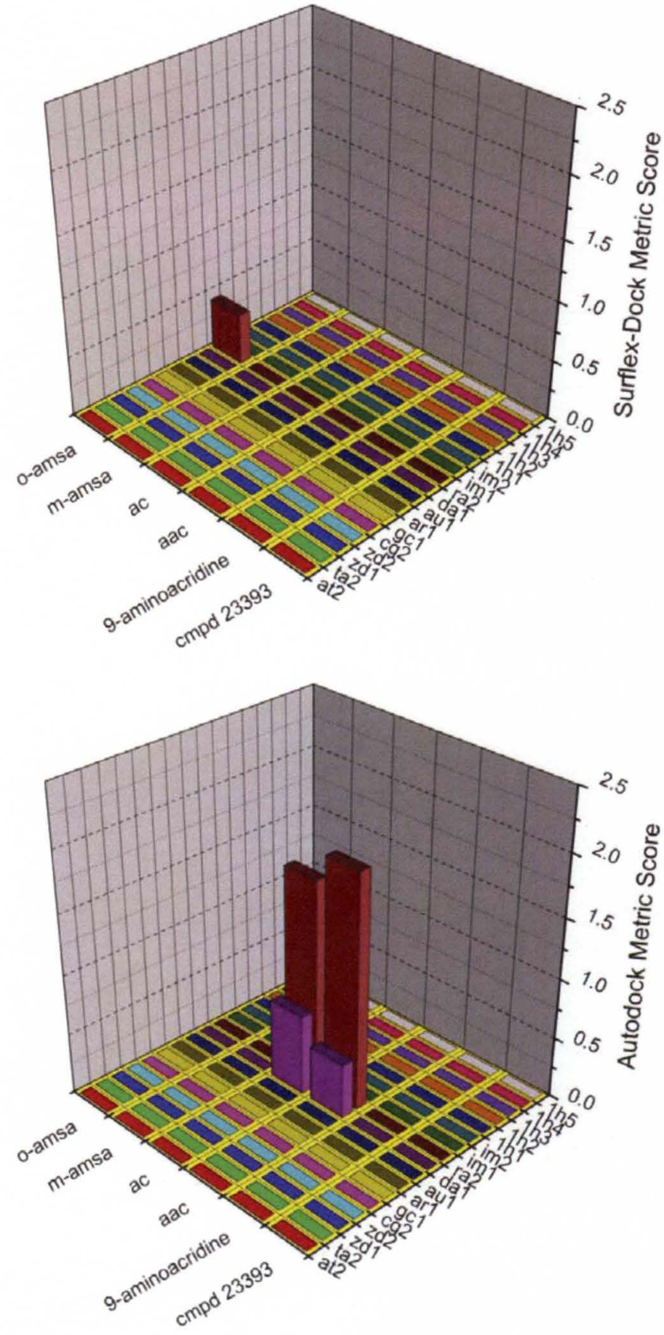
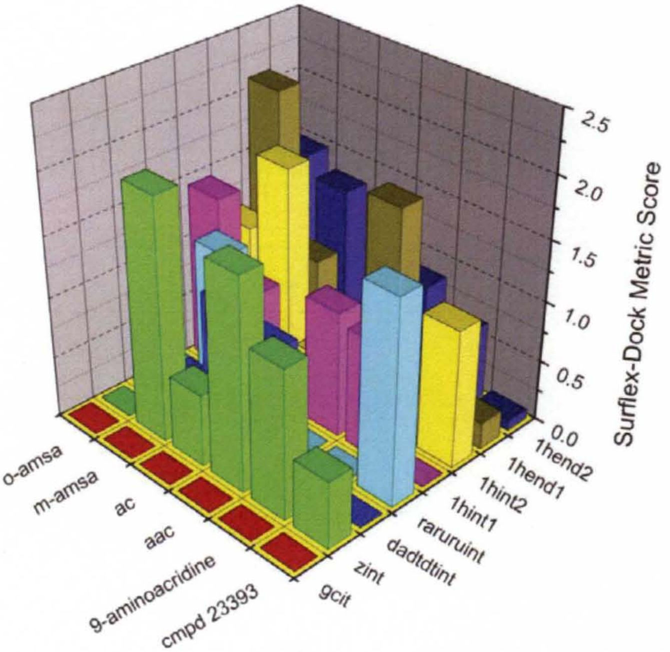


Figure 31. Surflex-Dock (top) and Autodock (bottom) docking scores for the acridine derivatives from the 67 Compound Set, <u>after</u> application of the <u>intercalation</u> metrics. The results shown are for the <u>intercalation</u> sites in the nucleic acid library.

Figure 31. Surflex-Dock (top) and Autodock (bottom) docking scores for the acridine derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.



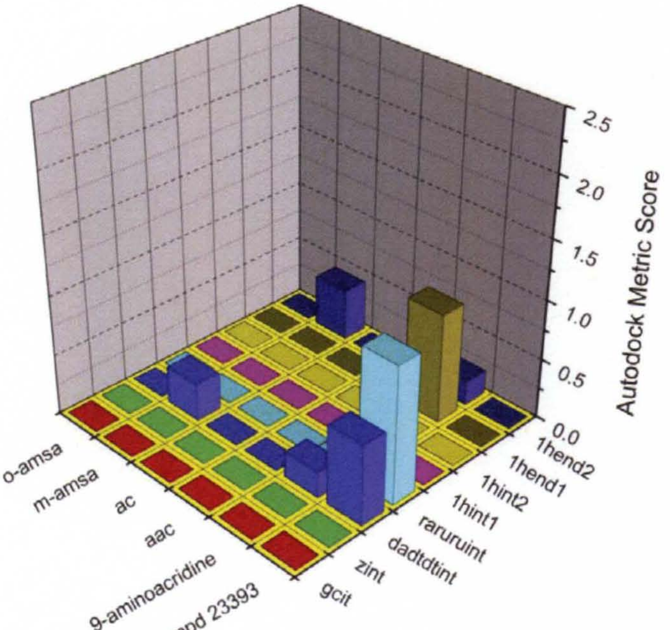
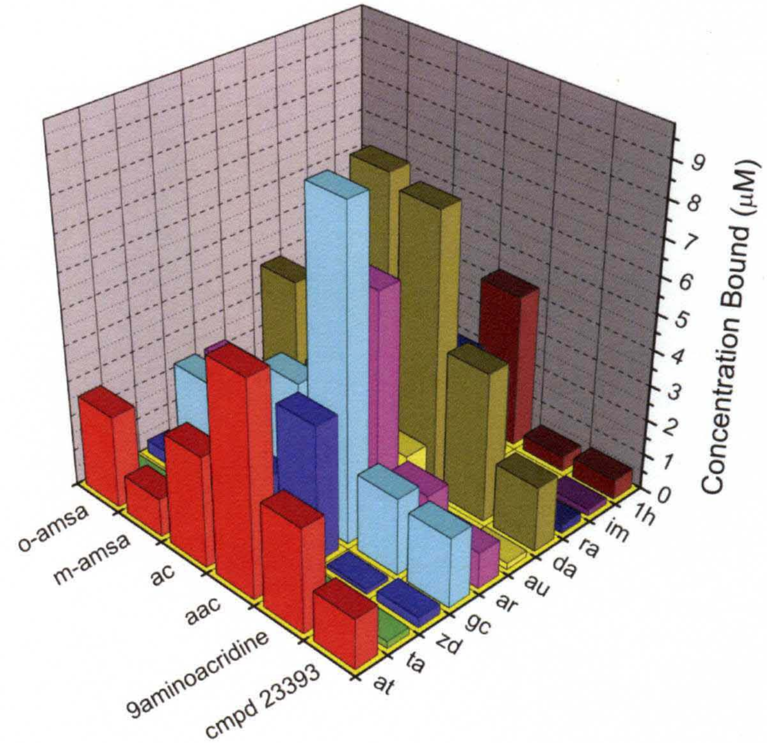


Figure 32. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the acridine derivative set from the 67 Compound Set

Figure 32. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the acridine derivative set from the 67 Compound Set



Aromatic Diamidine Derivatives. Members of the aromatic diamidine family (Figure 26) of compounds have proven to be effective treatments for many infectious diseases such as malaria and trypanosomiasis [111, 113]. These compounds have generally been shown to bind to AT rich DNA sequences and prefer binding to the DNA triplex, poly(dA)-$[\text{poly(dT)}]_2$, which is consistent with the competition dialysis data (Figure 33) [130]. The crescent shaped structure of many of the compounds initially suggests that the aromatic diamidines generally bind to the minor groove, and this is true for many molecules in this family [130]. The crescent shape assists with fitting the compound to the geometry of the minor groove and allows the aromatic diamidines to form hydrogen bonds at the base of the groove [113]. Interestingly, however, the position of the terminal imidazoline groups for all three of the compounds in our test data set (Figures 26) increases the planarity of the compounds and causes the preferred mode of binding to be intercalation into the triplex DNA structure [111, 130].

When the groove and intercalation metrics are applied to the aromatic diamidine derivative set, the scores show some positive scores for the grooves, but mostly positive scores for the intercalation sites (Figures 34 and 35). This is generally consistent with the intercalation mechanism of the aromatic diamidines described here. The observation that there are positive groove scores predicted by the *in silico* metrics is not entirely surprising. Subtle changes in the aromatic diamidines (for example the para-para to meta-meta switch of the terminal groups) can switch the main mode of binding from groove binding to intercalation, but the compounds still may possess secondary groove binding interactions. Thus, while the compounds listed here primarily intercalate, they could easily groove bind with only minimal structural changes. This emphasizes that the

Figure 33. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the aromatic diamidine derivative set from the 67 Compound Set.

Figure 33. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the aromatic diamidine derivative set from the 67 Compound Set.

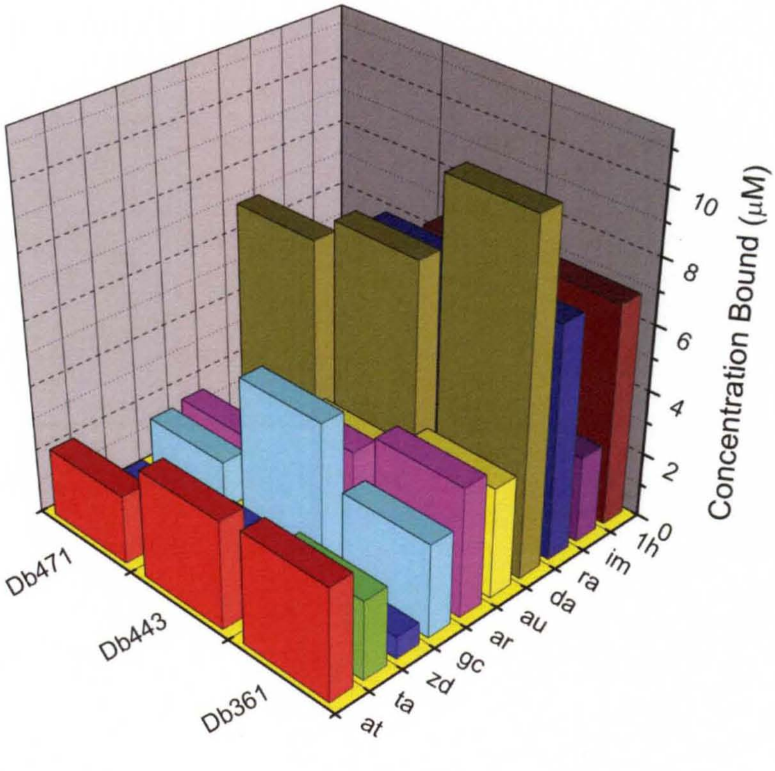
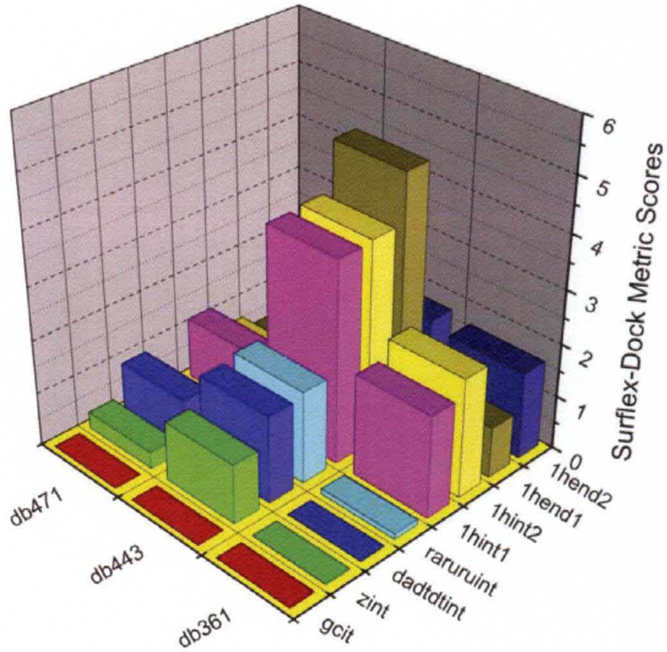


Figure 34. Surflex-Dock (top) and Autodock (bottom) docking scores for the aromatic diamidine derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.

Figure 34. Surflex-Dock (top) and Autodock (bottom) docking scores for the aromatic diamidine derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.



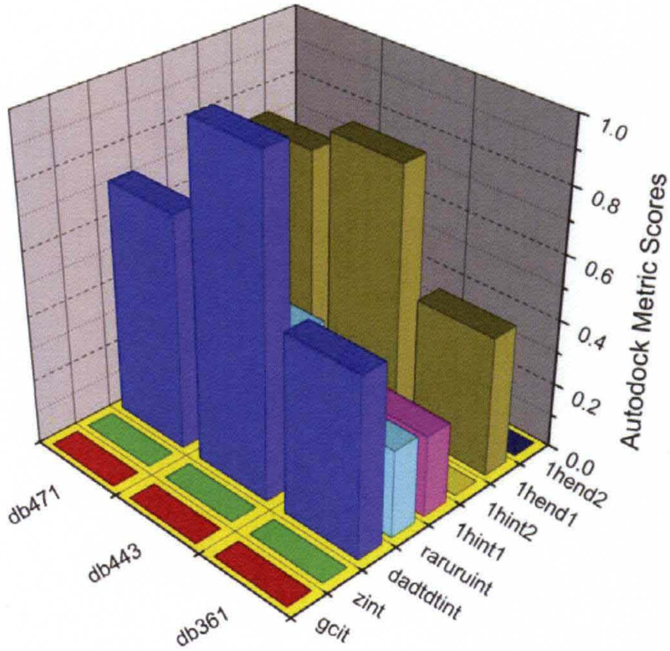
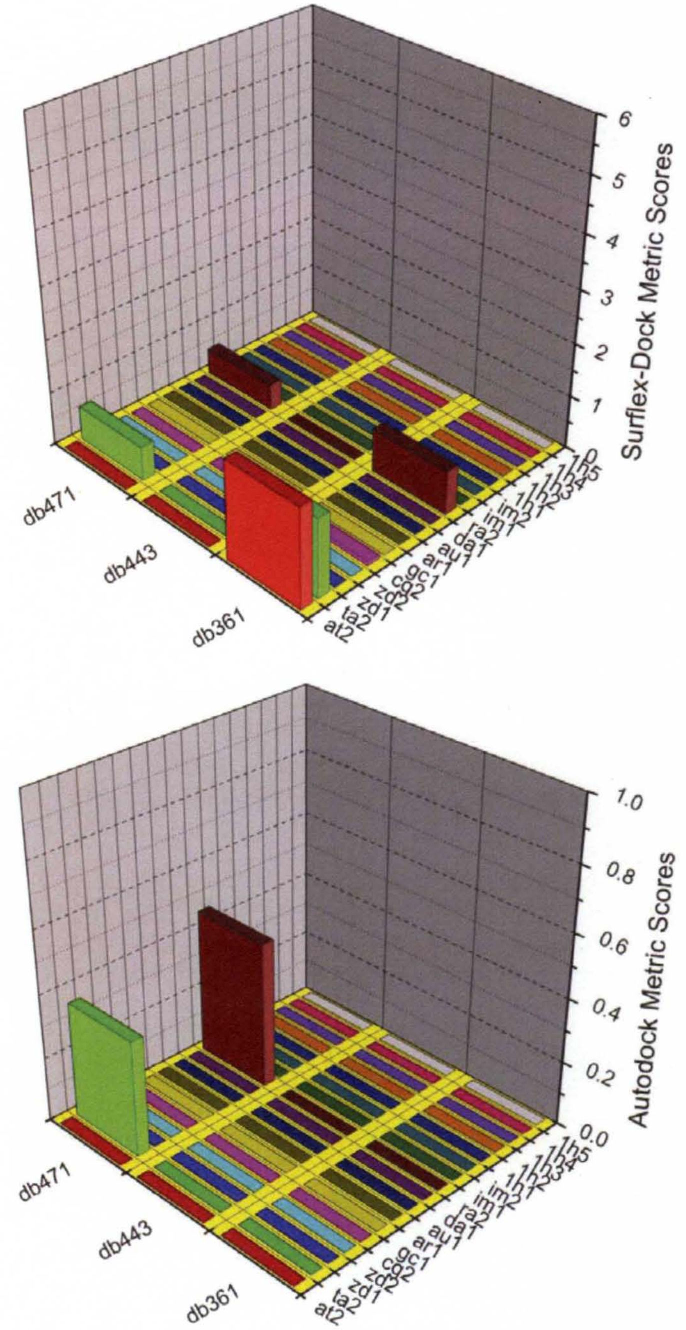


Figure 35. Surflex-Dock (top) and Autodock (bottom) docking scores for the aromatic diamidine derivatives from the 67 Compound Set, <u>after</u> application of the <u>groove binding</u> metrics. The results shown are for the <u>groove binding</u> sites in the nucleic acid library.

Figure 35. Surflex-Dock (top) and Autodock (bottom) docking scores for the aromatic diamidine derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.



experiments here are testing the limits of these software, as subtle chemical changes can make a significant difference in predicting the mode of binding, using the established metrics.

The competition dialysis data show that the aromatic diamidines generally bind triplex DNA and RNA with some binding to quadruplex DNA (Figure 33). The Surflex data predicts intercalation of these compounds mostly to quadruplex DNA and somewhat to the triplex DNA (Figure 34). Autodock predicts intercalation predominately into the triplex DNA, with minimal quadruplex intercalation (Figure 34). Much lower binding to the grooves is predicted, although the scores are most positive for AT duplex DNA and triplex RNA which is where groove binding of many aromatic diamidines occurs *in vitro* (Figure 35). The Autodock data in particular closely resembles the sequence specificity of the aromatic diamidines that was determined by competition dialysis. Overall, these data suggest that the metrics can generally elucidate the mode of binding of the aromatic diamidines as well as predict sequence specificity of these compounds.

Cyclic Aromatic Derivatives. This group of compounds, along with the "Other Compounds" set represents a diverse group of chemical compounds including porphyrins, the threading intercalator PIPER, and compounds that possess large, fused, aromatic chemical groups (Figure 26). As such, it is expected that these compounds should interact primarily by an intercalative or end-stacking mechanism. The porphyrins (including tetrakisporphrine and mesotetrakisporphine) are perhaps the best known class of G-quadruplex binding ligands. These compounds, in particular the compound TmPyP4, have been investigated in detail as their structure suggests that the compounds may preferentially stack and interact with the guanine quartet of quadruplex structures

[10, 131]. However, the porphyrins have also been shown to interact with the grooves of the human telomeric quadruplex, $AG_3(TTAGGG)_3$, through an "outside" groove binding mechanism [132]. Porphyrins also appear to generally suffer from non-specific binding to many other forms of duplex and triplex DNA and RNA, as has been demonstrated by competition dialysis [10]. On the other hand, the small molecule NMM, has been shown to be highly selective for G-quadruplexes over duplex and triplex nucleic acids, although NMM binds with lower affinity than porphyrins such as TmPyP4 [10]. The aromatic system of the bis acridine molecule Bisa, also reflects its propensity to intercalate into DNA as well as possibly act as a threading intercalator in various nucleic acids [133-134]. Finally, the last member of this family is PIPER which is reported to bind by end-stacking to various G-quadruplex nucleic acids and also possibly interacting by a threading intercalator mechanism [135].

The *in silico* Surflex data classifies all of these compounds primarily as intercalators, as there are no positive groove binding scores (Figure 36), but positive scores for several intercalation sites (Figure 37). The Autodock results are more diverse as the metrics predict some of the compounds to be exclusively groove binders (tetrakisporphine and bisa) while the others to act predominately by intercalation or endpasting (hoa, mesotetrakisporphine, nmm, piper). While it is likely that the porphyrin tetrakisporphine can interact with grooves, given its promiscuous binding behavior, it is well known that porphyrins in general can intercalate and endpaste into nucleic acids. Overall, Surflex appears superior at predicting the mechanism of action of these compounds compared to Autodock. For structure and sequence specificity, this group of compounds generally appears to bind with preference to triplex and quadruplex

Figure 36. Surflex-Dock (top) and Autodock (bottom) docking scores for the cyclic aromatic derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

Figure 36. Surflex-Dock (top) and Autodock (bottom) docking scores for the cyclic aromatic derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

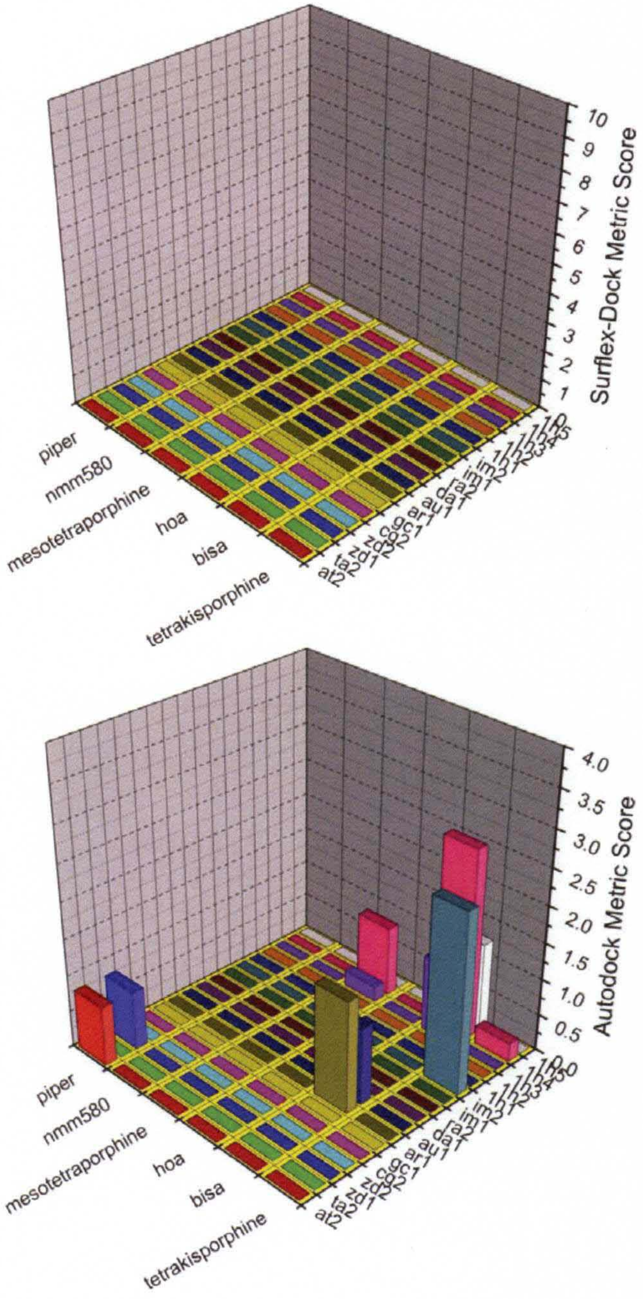
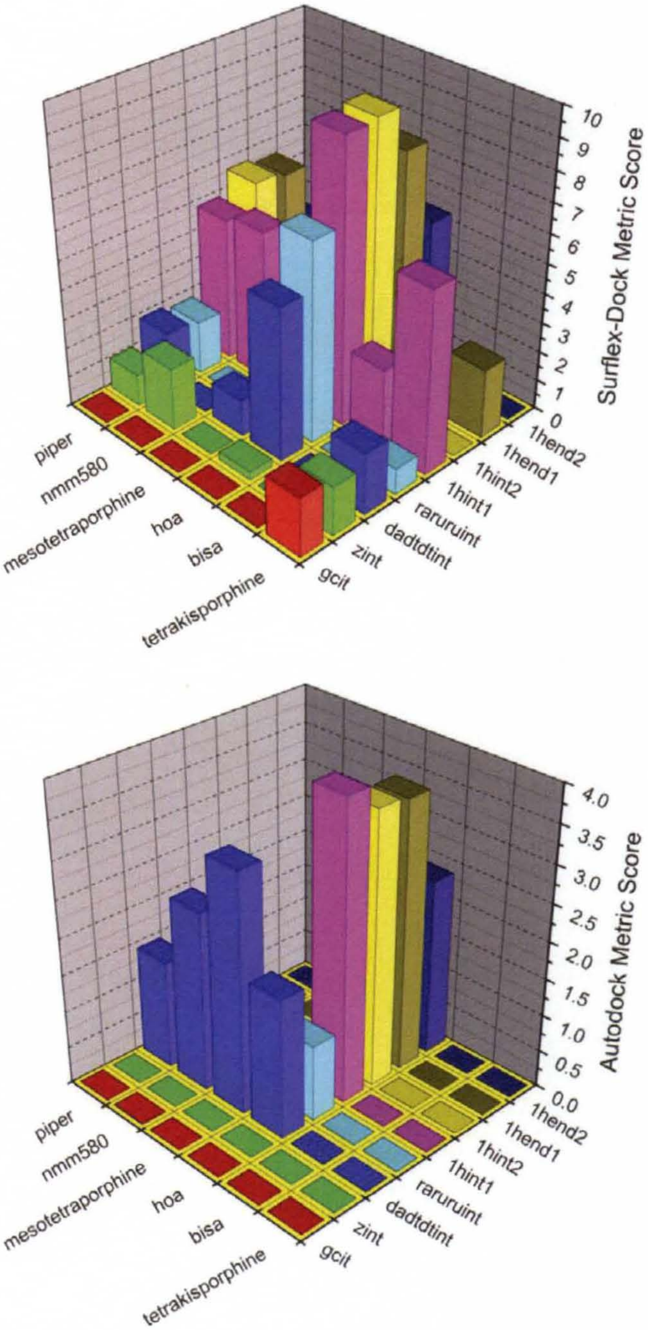


Figure 37. Surflex-Dock (top) and Autodock (bottom) docking scores for the cyclic aromatic derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.

Figure 37. Surflex-Dock (top) and Autodock (bottom) docking scores for the cyclic aromatic derivatives from the 67 Compound Set, <u>after</u> application of the <u>intercalation</u> metrics. The results shown are for the <u>intercalation</u> sites in the nucleic acid library.

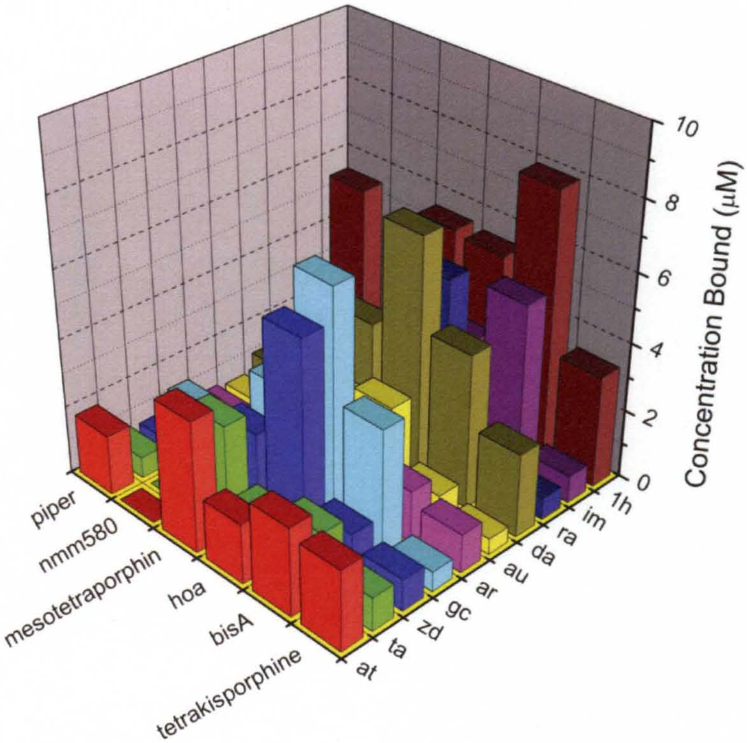


DNA, as determined by competition dialysis (Figure 38). The intercalator metrics from Surflex predict that these are also the preferred binding structures of DNA for all of the compounds (Figure 37). For the predicted intercalators using the Autodock metrics (hoa, mesotetrakisporphine, nmm, piper), these compounds are generally predicted to bind to the triplex DNA, except for hoa that is also predicted to bind to quadruplex DNA (Figure 37). In summary, for the cyclic aromatic set, Surflex appears superior to Autodock at predicting both the binding mechanism and sequence specificity of the compounds and the predicted data is in reasonable agreement with the data from competition dialysis.

Dibenzophenanthroline Derivatives. The dibenzophenanthrolines (Figure 26) were designed as small molecules that would intercalate into the triplex DNA poly(dA)-[poly(dT)]$_2$ [136]. The crescent-shaped curvature of the compounds suggests that the compounds may also bind to the grooves of DNA. These compounds posses a pentacyclic ring and are either monosubstituted (mpq2 and mpq3) or bisubstituted (mmq1, mmq3 and moq1). The monosubstitited compounds have reported preferential triplex binding compared to duplex, while the bisubstituted derivates show promiscuous binding to both duplex and triplex DNA, which is generally consistent with the competition dialysis results reported here. There are also recent reports of dibenzophenanthrolines suggesting that these molecules bind G-quadruplexes, as the fused aromatic chemical features provide large π–orbital stacking with the guanine tetrads of the G-quadruplexes [123, 136-137].

Figure 38. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the cyclic aromatic derivative set from the 67 Compound Set.

Figure 38. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the cyclic aromatic derivative set from the 67 Compound Set.



The Surflex *in silico* docking results appear to generally segregate the dibenzophenanthrolines on the basis of whether they are monosubstituted (mpq2 and mpq3) or bisubstituted (mmq1, mmq3, moq1) (Figures 39 and 40). The monosubstituted ligands have few positive groove binding scores and instead have mostly positive intercalation site scores, suggesting intercalation is the mechanism of action. On the other hand, the bisubstituted ligands appear to be classified largely as groove binders according to the groove binding rules developed with Surflex (Figure 39). This is consistent with the more crescent shaped curvature of the bisubstituted ligands which have substituent locations that would support groove binding. Surflex appears to be able to modestly elucidate the sequence specificity of the classified mono and bisubstituted compounds. The groove binding scores that are positive are generally for the duplex AT rich DNA which the bisubstituted ligands are known to bind, as determined by competition dialysis (Figure 41). The intercalation binding scores that are the most positive are typically for the triplex and quadruplex sites for the monosubstituted compounds. The Autodock data are less clear, as there are positive scores present for the grooves and intercalation sites in both the mono and bisubstituted ligands, suggesting that the Autodock metrics are less successful at classifying the compounds as primarily groove binders or intercalators (Figure 39 and 40). However, Autodock does appear to be able to identify the sequence preference for the general class of compounds, as the highest groove binding scores are found for AT rich duplex DNA and the triplex nucleic acids (Figure 39). Also, the highest Autodock intercalation scores are found for the triplex intercalation sites, which were the basis for the original design of these small molecules (Figure 40). In summary, the metrics have only moderate success at

Figure 39. Surflex-Dock (top) and Autodock (bottom) docking scores for the dibenzophenanthroline derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

Figure 39. Surflex-Dock (top) and Autodock (bottom) docking scores for the dibenzophenanthroline derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

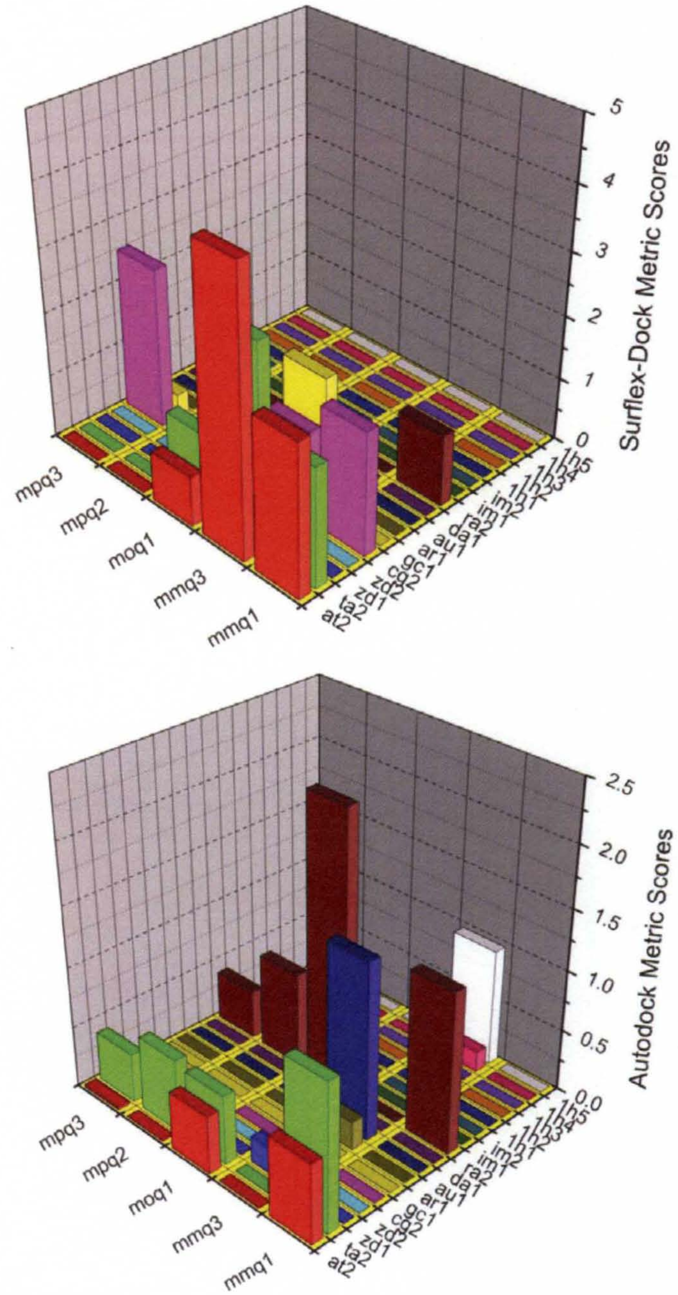
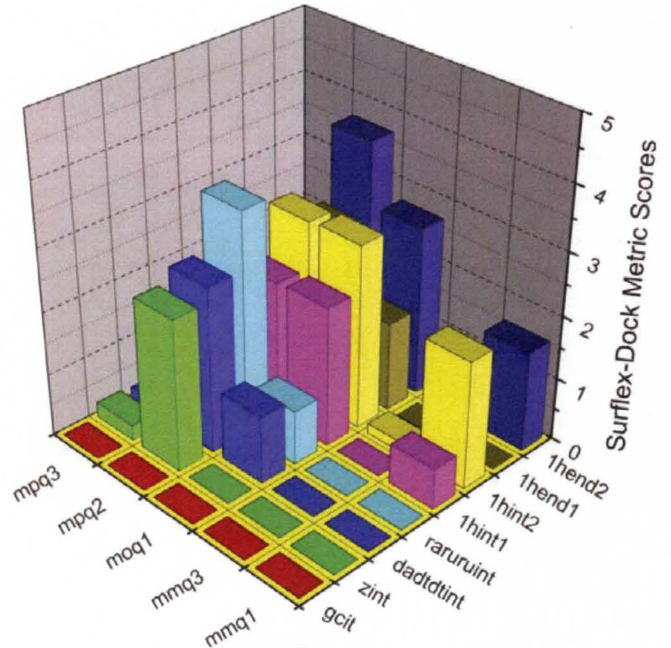


Figure 40. Surflex-Dock (top) and Autodock (bottom) docking scores for the dibenzophenanthroline derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.

Figure 40. Surflex-Dock (top) and Autodock (bottom) docking scores for the dibenzophenanthroline derivatives from the 67 Compound Set, <u>after</u> application of the <u>intercalation</u> metrics. The results shown are for the <u>intercalation</u> sites in the nucleic acid library.



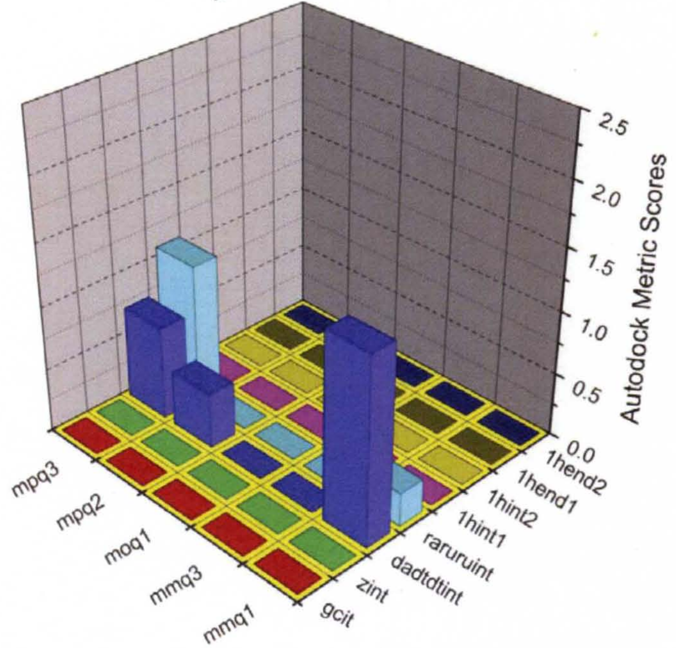
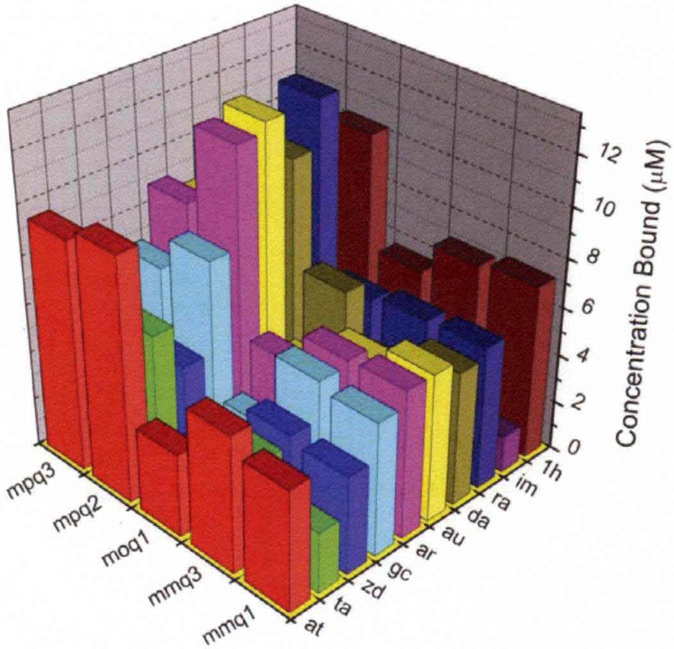


Figure 41. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the dibenzophenanthroline derivative set from the 67 Compound Set.

Figure 41. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the dibenzophenanthroline derivative set from the 67 Compound Set.



predicting the mechanism of action and sequence selectivity of the dibenzophenanthroline derivatives. This may be ascribed to the minor chemical changes among these compounds that can result in a change in the binding mode.

Bis-quinoline Derivatives. There is interest in "bis" intercalators as these compounds can intercalate into two sites in nucleic acids which allows increased affinity and specificity of the small molecule for the nucleic acid [121]. Previous results have shown that some of these compounds preferentially intercalate into triplex and quadruplex DNA structures [123]. These compounds are unique as they have a long linking chain that connects the two quinoline derivatives (Figure 26). This chain is capable of binding to the groove of the nucleic acid and thus these compounds exhibit both intercalation and groove binding character that may challenge the metrics as described here.

Application of the *in silico* Surflex metrics shows that positive scores are present in the groove sites and the intercalation sites, suggesting that these compounds have substantial groove binding and intercalation character (Figures 42 and 43). This is possible given that the planar, aromatic groups intercalate into DNA and the long linker likely binds in the groove. The most positive scores for the groove sites are with AT rich DNA, an observation that is generally consistent with the known competition dialysis data which shows binding to AT rich DNA (Figure 44). The most positive intercalation scores are for the quadruplex nucleic acids which these ligands are known to interact. Autodock classifies these ligands exclusively as groove binders as there are only positive groove binding scores present (Figure 42). This could be due to the particularly long linker chains between the quinoline groups, which can lie in the groove and dramatically impact Autodock scores.

Figure 42. Surflex-Dock (top) and Autodock (bottom) docking scores for the bis-quinoline derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

Figure 42. Surflex-Dock (top) and Autodock (bottom) docking scores for the bis-quinoline derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

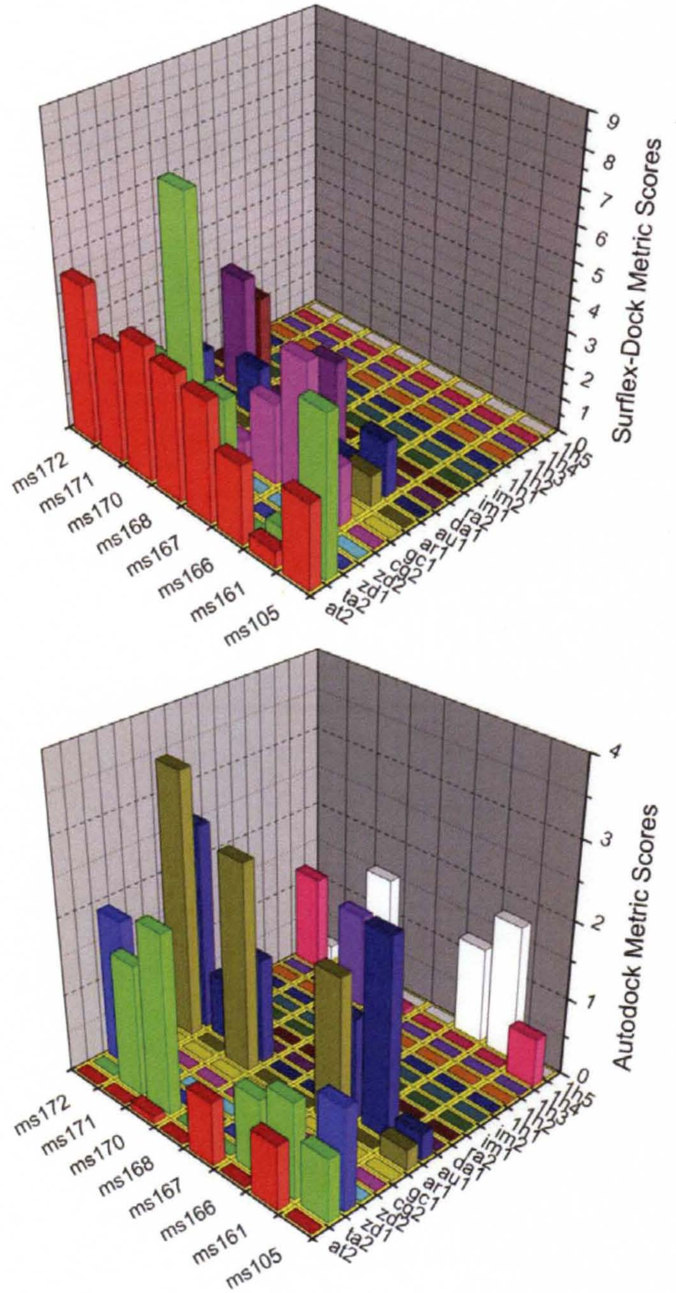
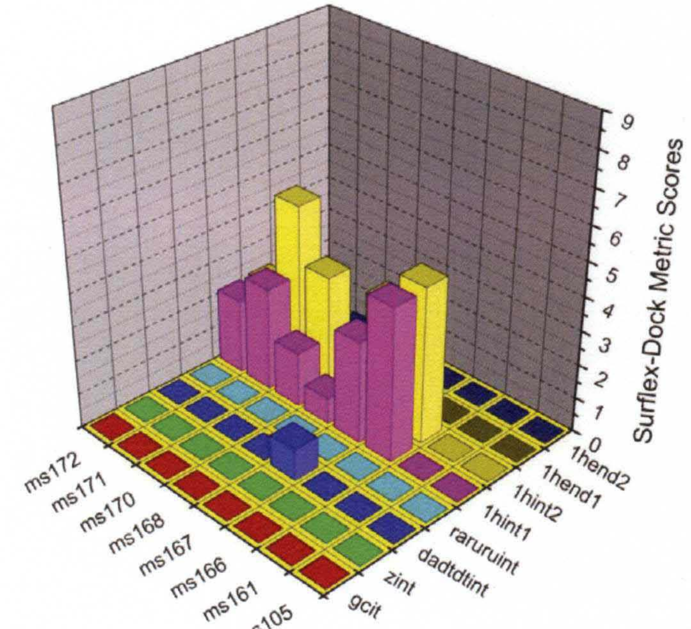


Figure 43. Surflex-Dock (top) and Autodock (bottom) docking scores for the bis-quinoline derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.

Figure 43. Surflex-Dock (top) and Autodock (bottom) docking scores for the bis-quinoline derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.



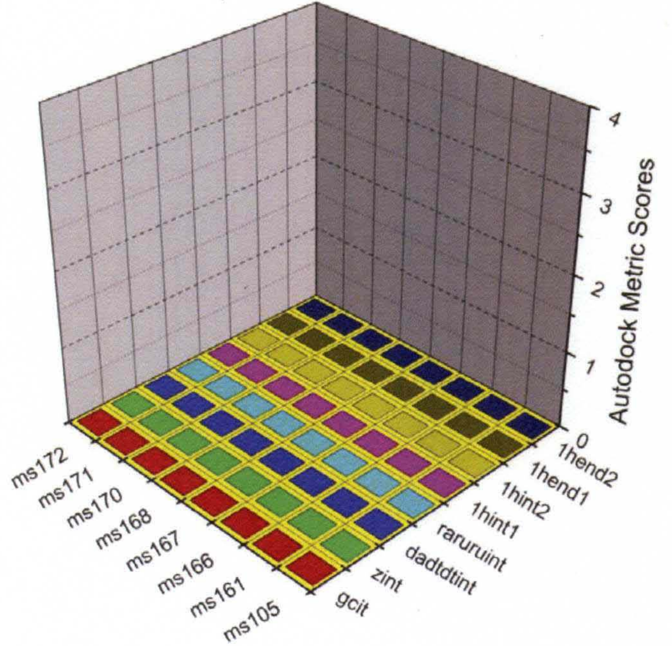
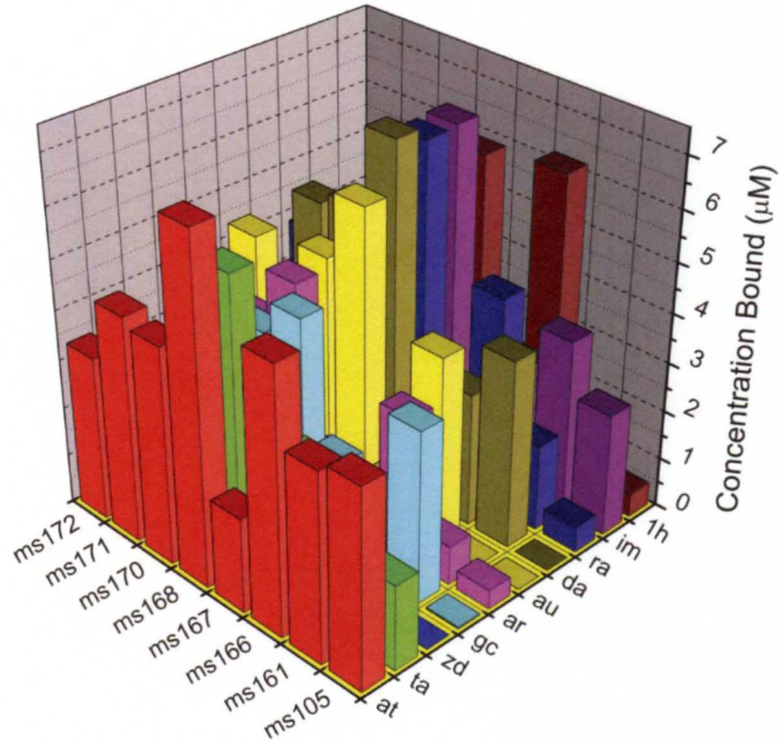


Figure 44. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the bis-quinoline derivative set from the 67 Compound Set.

Figure 44. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the bis-quinoline derivative set from the 67 Compound Set.



Interestingly, the Autodock results do appear to generally mimic the sequence specificity of the compounds, as positive groove scores are seen for the AT rich DNA, the RNA and the quadruplex structures. Overall, however, Surflex again appears best at predicting the mechanism of action and sequence specificity of the Bis-quinoline derivatives. It is worth noting for this class of compound that the size and extended length of these molecules make these one of the most challenging docking experiments of all of the compound sets tested.

Amidoanthraquinones (aromatic side chains and non-aromatic side chains). The amidoanthraquinones (Figure 26) have been reported to bind to various nucleic acids depending on the location of the side chains that extend from the central fused aromatic ring system. The so-called 1,4-disubstituted small molecules (tcj74 and tcj62) appear to bind duplex DNA while the 2,6-disubstituted small molecules (tcj78, telominh1 and telominh5 and tcj69) and the 2,7-disubstituted small molecules (pjp57 and pjp66) prefer triplex over duplex DNA [123, 138-139]. Additionally, the amidoanthraquinones have been shown to bind G-quadruplex nucleic acids, as is confirmed by the competition dialysis data (Figure 45) [140-141]. These compounds have a primary reported intercalation mechanism of binding with additional "threading" behavior, where the fused aromatic system intercalates, but the substituents can occupy the grooves of the nucleic acids [123, 138, 140].

For the amidoanthraquinones with aromatic side chains, the *in silico* groove binding metrics for Surflex show positive scores with the AT rich duplex DNA (Figure 46). However, the scores for the intercalation sites after applying the metrics are generally higher and suggest that intercalation is the primary mechanism of action, with

Figure 45. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the amidoanthraquinones (aromatic side chains) derivative set from the 67 Compound Set.

Figure 45. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the amidoanthraquinones (aromatic side chains) derivative set from the 67 Compound Set.

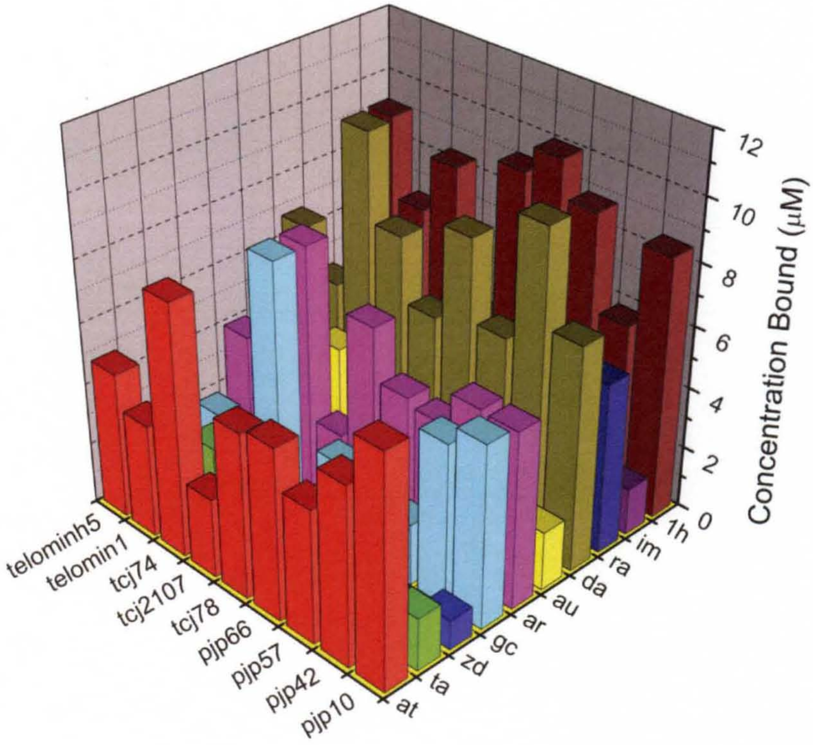
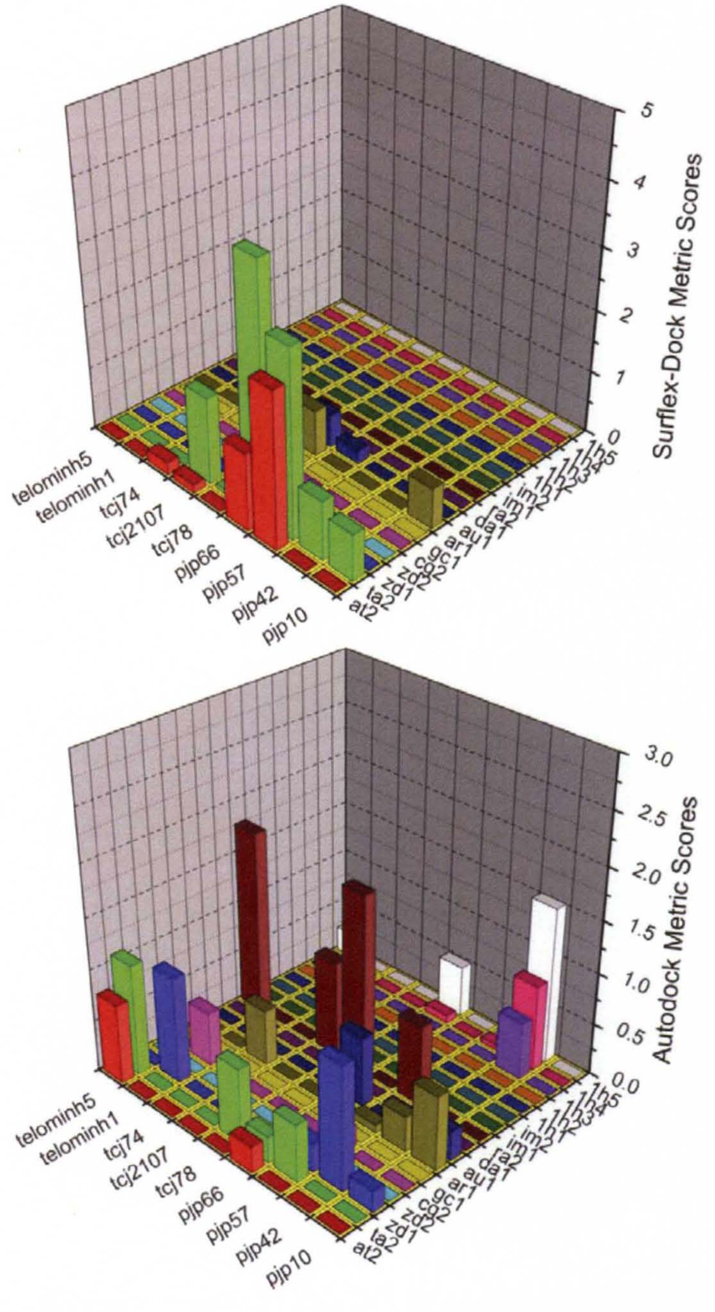


Figure 46. Surflex-Dock (top) and Autodock (bottom) docking scores for the amidoanthraquinones (aromatic side chains) derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

Figure 46. Surflex-Dock (top) and Autodock (bottom) docking scores for the amidoanthraquinones (aromatic side chains) derivatives from the 67 Compound Set, <u>after</u> application of the <u>groove binding</u> metrics. The results shown are for the <u>groove binding</u> sites in the nucleic acid library.

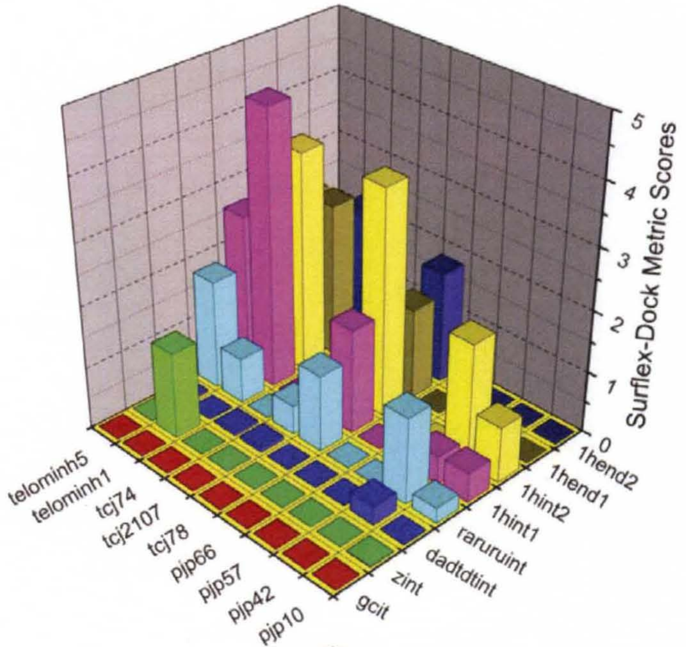


some additional groove binding interactions possible (Figure 47). This is consistent with the known binding mechanism of action of many of these compounds. Interestingly, the highest *in silico* scores are found typically for the triplex and quadruplex intercalation sites, suggesting that these structures are the preferred intercalation sites, which is generally consistent with their known structural preference. The Autodock groove binding metrics yield less clear information, as there appears to be positive scores for many ligands to a number of different grooves (Figure 46). Application of the intercalation metrics for Autodock shows that many of the amidoanthraquinones actually appear to prefer to intercalate into the triplex DNA (Figure 47). However, no positive scores are seen for quadruplex intercalation sites which is somewhat surprising given that these compounds are generally known to bind quadruplexes. Generally, the score distribution for the groove and intercalation sites for Autodock support intercalation as the primary mechanism of action, with some groove binding behavior evident.

For the amidoanthraquinones with non-aromatic side chains, the Surflex metrics clearly predict intercalation as the preferred mechanism of action, with some groove binding possible (Figures 48 and 49). Similar to the results seen with the amidoanthraquinones with aromatic side chains, both the triplex and quadruplex intercalation sites have the highest scores, suggesting that these are the preferred structures, which is generally consistent with the competition dialysis data (Figure 50). Application of the Autodock metrics classifies these compounds as almost exclusively groove binding in nature, as almost all of the positive scores seen are in the groove binding sites as opposed to the intercalation sites (Figures 48 and 49). We believe this may be due to the nature of the side chain substituents. In this class of compounds, the side chains are non-aromatic

Figure 47. Surflex-Dock (top) and Autodock (bottom) docking scores for the amidoanthraquinones (aromatic side chains) derivatives from the 67 Compound Set, <u>after</u> application of the <u>intercalation</u> metrics. The results shown are for the <u>intercalation</u> sites in the nucleic acid library.

Figure 47. Surflex-Dock (top) and Autodock (bottom) docking scores for the amidoanthraquinones (aromatic side chains) derivatives from the 67 Compound Set, <u>after</u> application of the <u>intercalation</u> metrics. The results shown are for the <u>intercalation</u> sites in the nucleic acid library.



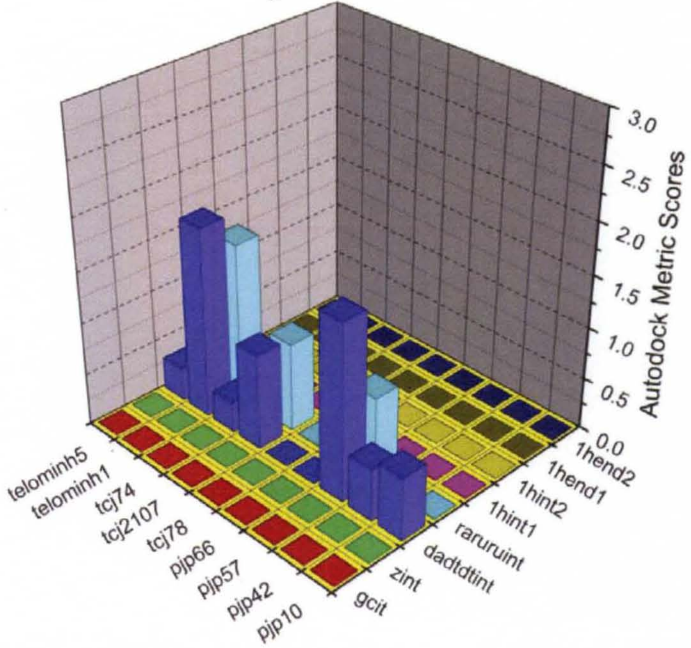


Figure 48. Surflex-Dock (top) and Autodock (bottom) docking scores for the amidoanthraquinones (non-aromatic side chains) derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

Figure 48. Surflex-Dock (top) and Autodock (bottom) docking scores for the amidoanthraquinones (non-aromatic side chains) derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

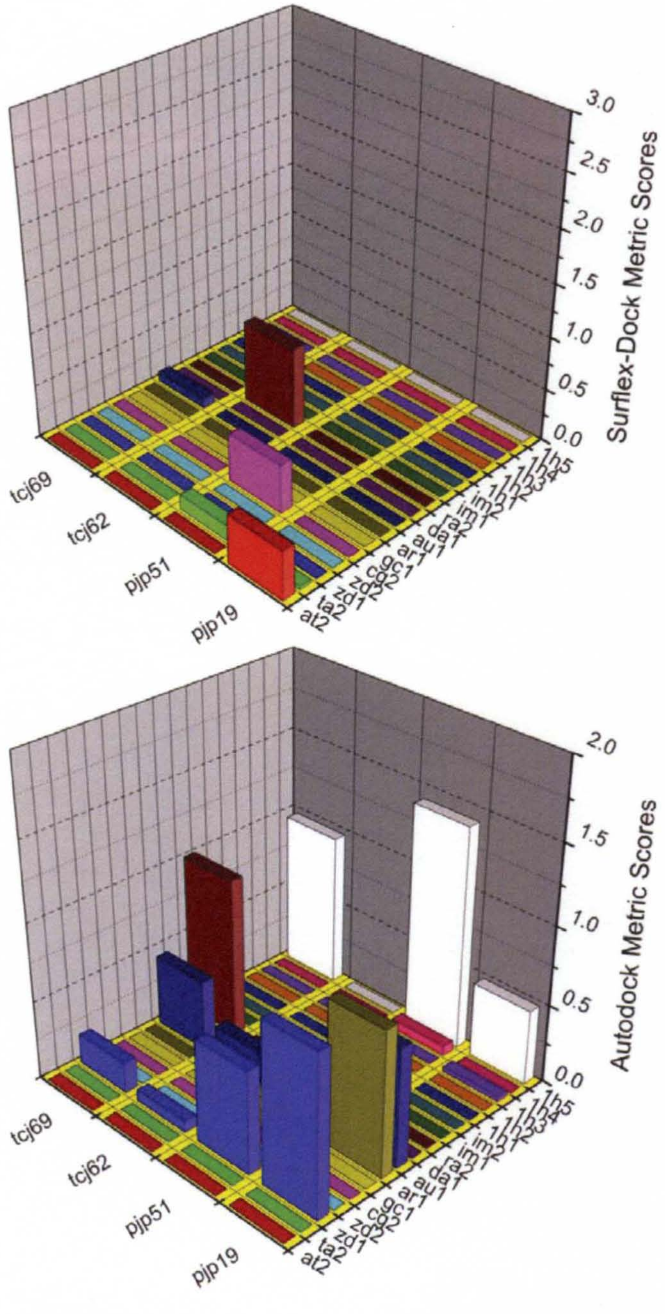


Figure 49. Surflex-Dock (top) and Autodock (bottom) docking scores for the amidoanthraquinones (non-aromatic side chains) derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.

Figure 49. Surflex-Dock (top) and Autodock (bottom) docking scores for the amidoanthraquinones (non-aromatic side chains) derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.

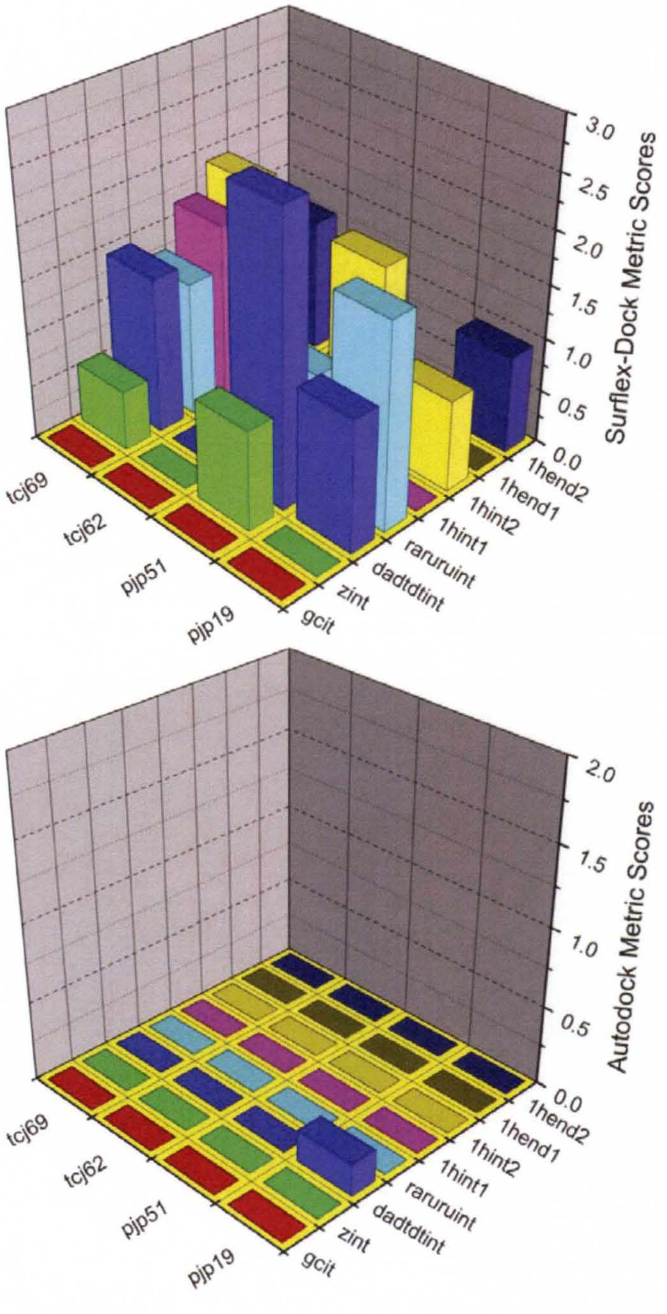
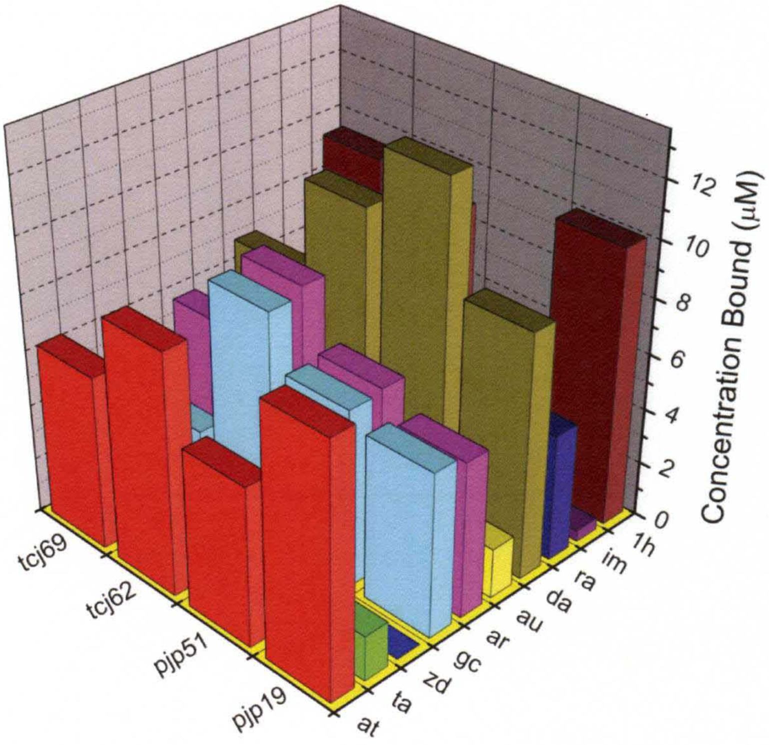


Figure 50. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the amidoanthraquinones (non-aromatic side chains) derivative set from the 67 Compound Set.

Figure 50. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the amidoanthraquinones (non-aromatic side chains) derivative set from the 67 Compound Set.



carbon chains, which should actually have superior groove binding properties as compared to the aromatic side chains. Thus, the Autodock metrics appear to score these compounds as having better groove binding character, which is likely consistent with their binding properties, but comes at the expense of classifying the compounds as predominately groove binders instead of intercalators. Overall, the results suggest that Surflex and Autodock are generally able to predict the binding mechanism of action of these compounds but prediction of sequence and structural specificity is more challenging.

Naphthoflavones. The flavonoids class of compounds have long been known to exhibit anti-inflammatory and antiviral properties (Figure 26) [142]. The alpha and beta naphthoflavone flavonoids included and tested here are known to intercalate into triplex DNA with high specificity, with little or no perceived binding to other nucleic acid structures [123, 142]. This binding behavior has also been seen in the competition dialysis data that was acquired on these compounds (Figure 51).

The Surflex and Autodock metric data for this class of compounds is perhaps the best out of all of the classes of compounds that were tested *in silico* with respect to differentiating groove binders versus intercalators. There are no positive scores present for either Surflex or Autodock with respect to the groove binding sites (Figure 52). Surflex has the most positive scores in the quadruplex intercalation sites while Autodock has the most positive scores in the triplex intercalation sites (Figure 53). This is important in several respects. First, both software predict exclusive intercalation of these compounds, with no discernable groove binding. This is entirely consistent with the reported literature and is expected given the planar, aromatic structure of the

Figure 51. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the naphthoflavone derivative set from the 67 Compound Set.

Figure 51. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the naphthoflavone derivative set from the 67 Compound Set.

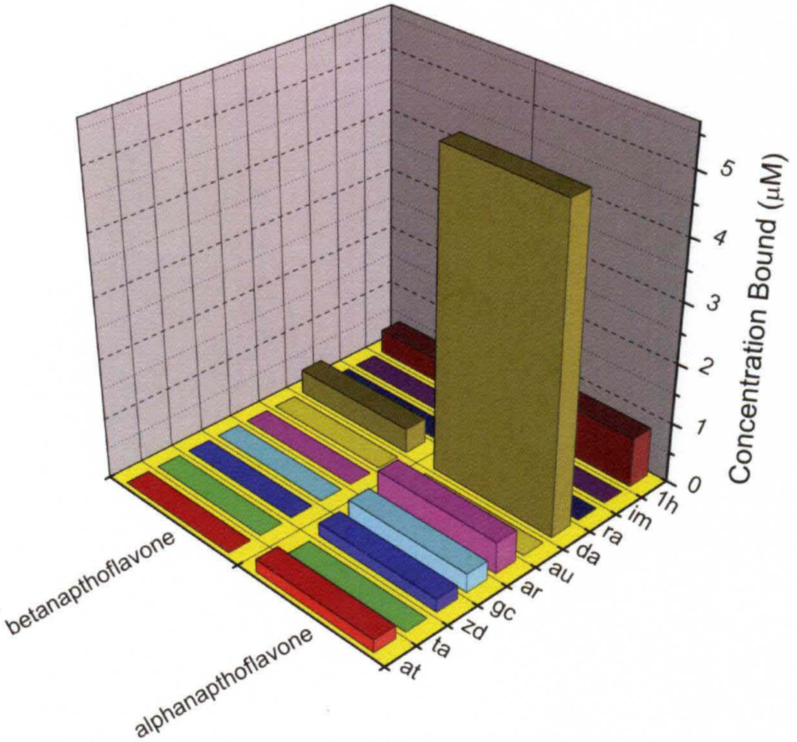


Figure 52. Surflex-Dock (top) and Autodock (bottom) docking scores for the naphthoflavone derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

Figure 52. Surflex-Dock (top) and Autodock (bottom) docking scores for the naphthoflavone derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

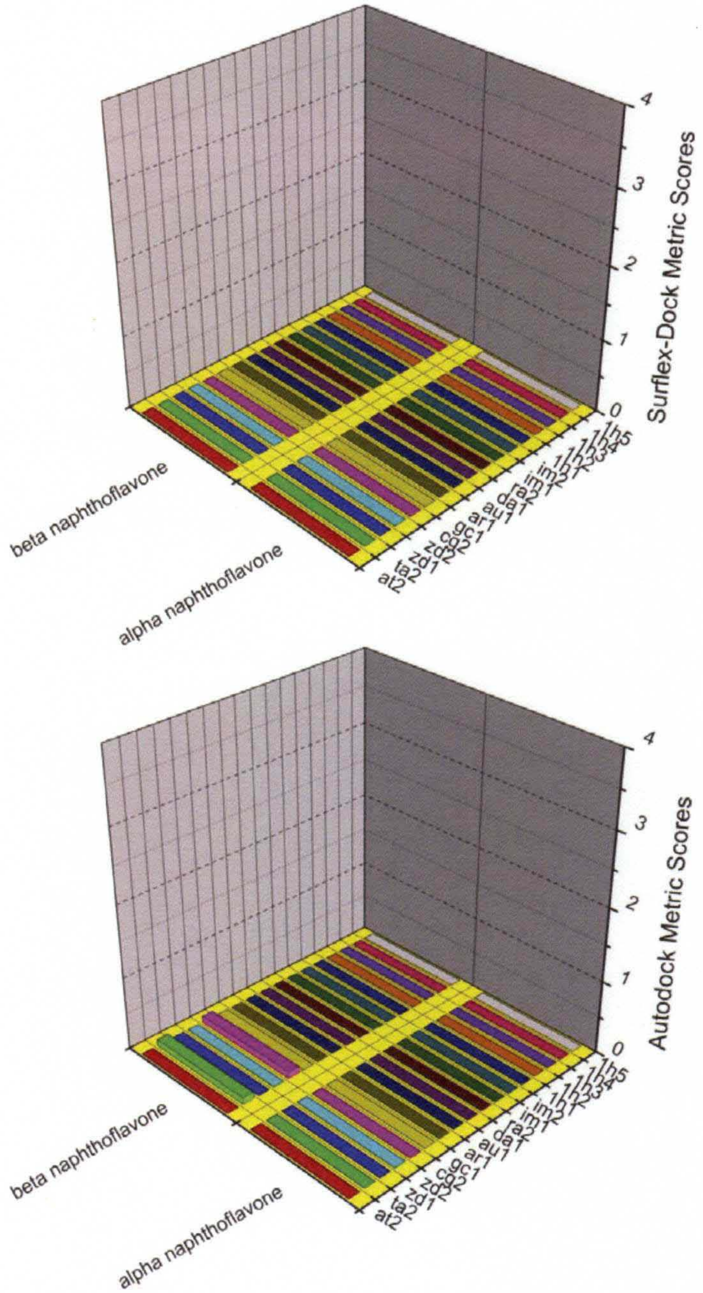
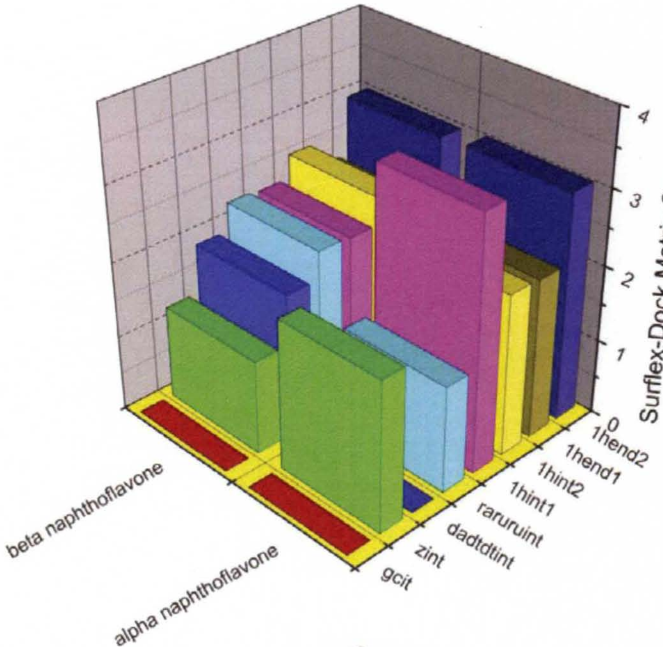
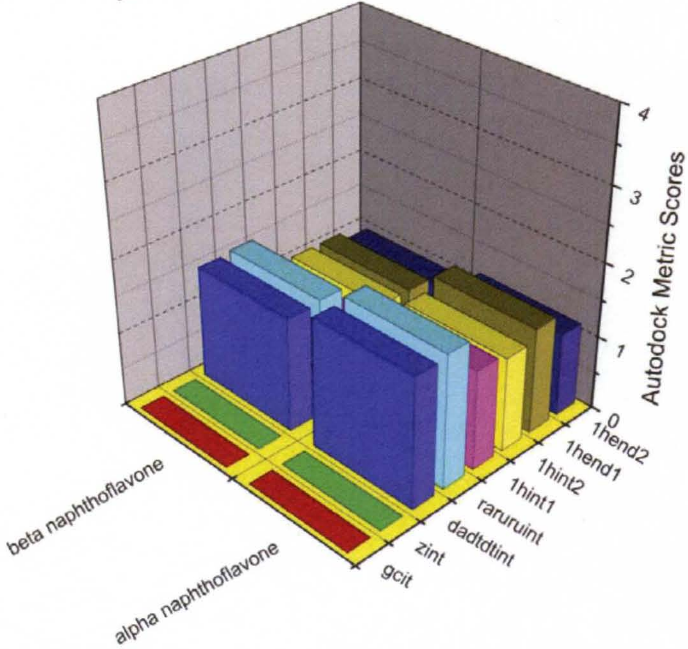


Figure 53. Surflex-Dock (top) and Autodock (bottom) docking scores for the naphthoflavone derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.

Figure 53. Surflex-Dock (top) and Autodock (bottom) docking scores for the naphthoflavone derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.





naphthoflavones. Second, this is an example of where Autodock appears superior to Surflex as Autodock predicts preferential binding to triplex DNA which is what occurs *in vitro* while Surflex predicts binding to quadruplex structures. The success in this class of compounds may also be due to their small size and few rotatable bonds. Smaller compounds with fewer rotatable bonds are typically much easier to dock compared to larger molecules because of the fewer degrees of freedom of the smaller compounds [63]. These results suggest that the metrics for both Surflex and Autodock have potential for successfully classifying ligands based on mechanism of action and structural preference.

Amidofluorenone Derivatives. The synthesis of the amidofluorenones (Figure 26) came largely out of the observed success of the anthraquinones at binding to G-quadruplexes and inhibiting the enzyme telomerase. The fluorenones were designed with the goal of achieving similar inhibitory potencies but with fewer cytotoxic side effect of the anthraquinones [143]. Modeling studies have suggested that these compounds can bind to nucleic acids by intercalation or end-stacking [143-144]. The side chains also suggest some groove binding occurs, imparting a "threading" type of intercalation binding behavior to these compounds.

Both the Surflex and Autodock metrics appear to have some positive scores particularly for the triplex nucleic acid grooves, suggesting that the amidofluorenones have groove binding character (Figure 54). However, the majority of the positive scores for Surflex are in the intercalation sites, supporting intercalation as the primary mechanism of action, with some groove binding behavior present (Figure 55). The *in silico* Surflex data generally predicts that intercalation to quadruplex structures is

Figure 54. Surflex-Dock (top) and Autodock (bottom) docking scores for the amidofluorenone derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

Figure 54. Surflex-Dock (top) and Autodock (bottom) docking scores for the amidofluorenone derivatives from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

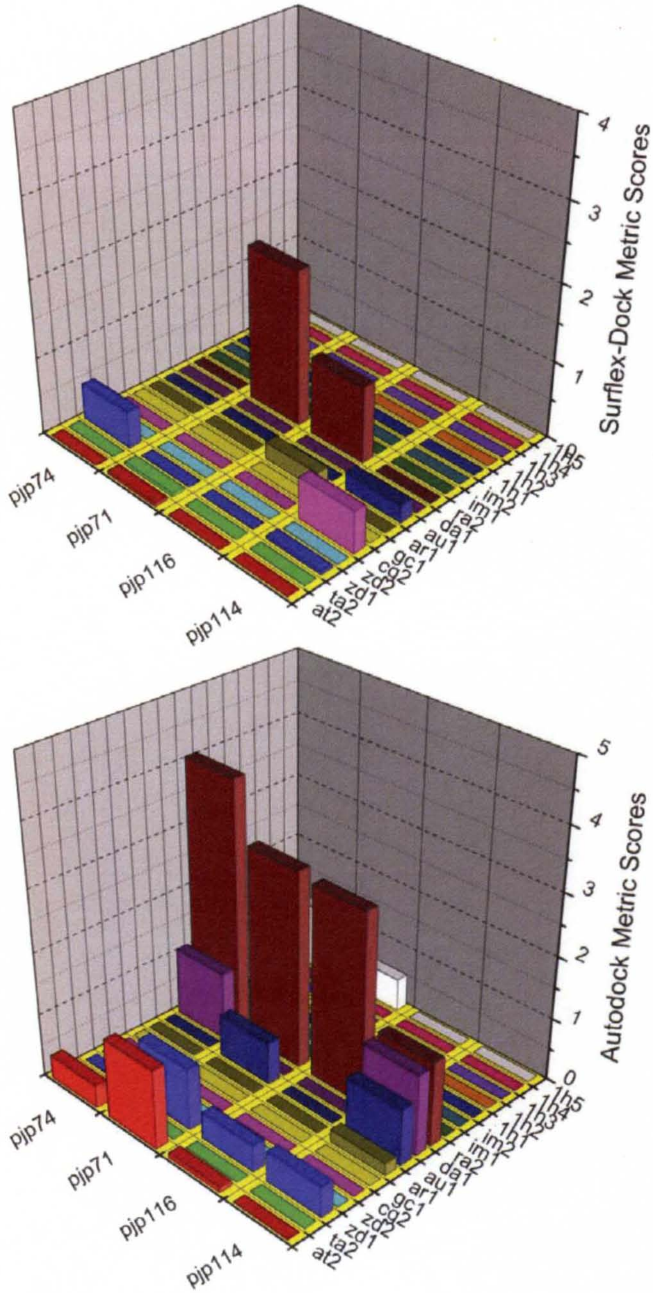


Figure 55. Surflex-Dock (top) and Autodock (bottom) docking scores for the amidofluorenone derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.

Figure 55. Surflex-Dock (top) and Autodock (bottom) docking scores for the amidofluorenone derivatives from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.

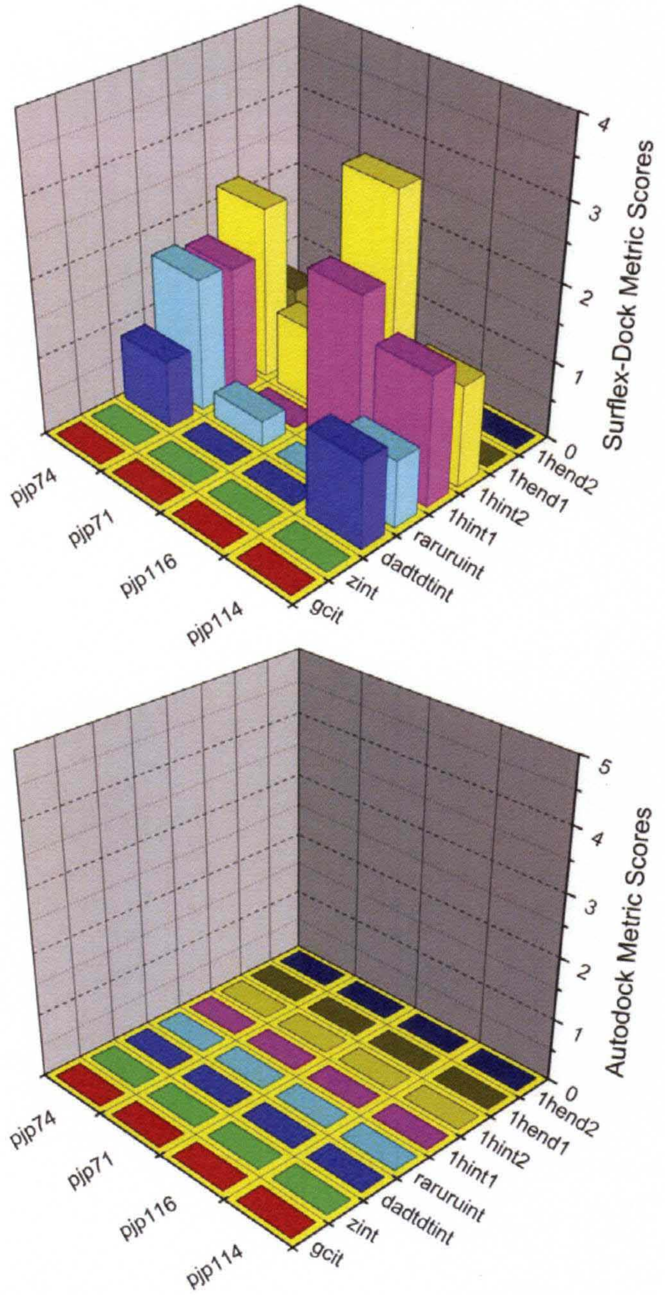
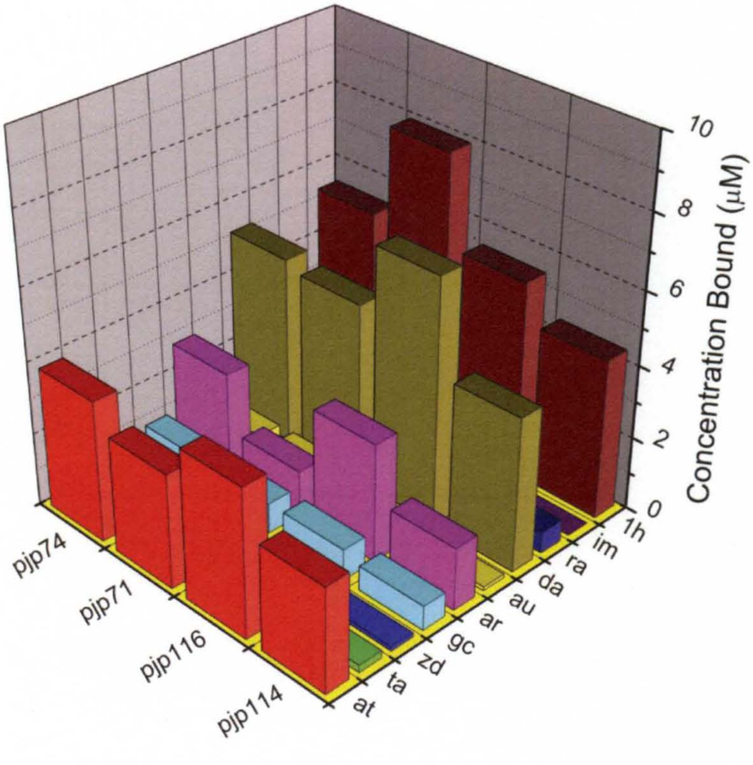


Figure 56. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the amidofluorenone derivative set from the 67 Compound Set.

Figure 56. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the amidofluorenone derivative set from the 67 Compound Set.

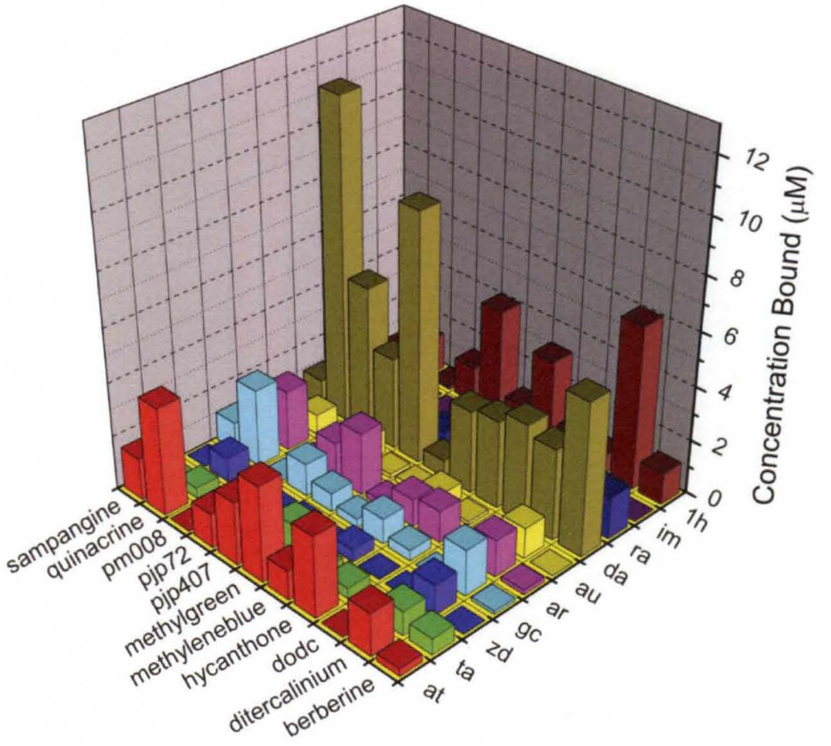


preferred over triplex, while the competition dialysis data generally shows comparable binding of these compounds to both triplex and quadruplex structures (Figure 56). In contrast, the Autodock data suggests that these compounds bind exclusively by groove binding, as there are only positive scores present for the groove sites (Figure 54) and no positive scores for the intercalation sites (Figure 55). The Autodock results are not overly surprising given that similar problems were observed with the structurally related amidoanthraquinones. For this class of compounds, Surflex appears superior to Autodock at predicting the mechanism of binding as well as structural specificity.

Other Compounds. Compounds that did not fall into any other chemical group have been included in the "Other Compounds" category and possess a large amount of structural diversity and nucleic acid binding specificity (Figure 26). The known mechanism of binding and sequence selectivity of these compounds is described briefly here. The competition dialysis data suggests that many of these compounds appear to favor triplex DNA, as well as interactions with quadruplex DNA and AT B-DNA to a lesser extent (Figure 57). Berberine has been shown to bind to predominantly triplex and quadruplex nucleic acids by intercalation or end-stacking [112, 145-147]. A preference for AT base pairs is notable for berberine [147]. Ditercalinium acts as a bis-intercalator with the linker sequence binding in the major groove of duplex DNA [148-149]. The interactions with the major groove are noteworthy as most small molecules interact with the minor groove [149]. Additionally, there have been reports of ditercalinium favoring Guanine-Cytosine over Adenine-Thymine sequences [149]. DODC has been identified as preferentially binding AT rich triplex DNA and to a lesser extent, quadruplex DNA structures [10, 39, 150]. The ligand appears to interact with different grooves of different

Figure 57. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the Other Compound set from the 67 Compound Set.

Figure 57. Competition Dialysis data showing the concentration of ligand bound to each nucleic acid structure for the Other Compound set from the 67 Compound Set.



quadruplexes [150]. Less significant interactions have been reported with minor groove interactions in duplex DNA [150]. Hycanthone is recognized as an intercalator that appears to prefer AT sequences over GC sequences [151]. This molecule is a particularly interesting "non-classical" intercalator, as it lacks a charge on the cyclic ring [151]. Methylene blue and Methyl green are unique compounds and have been included here as they may interact with the major groove of DNA. These compounds may be inappropriate for classification based on the metrics developed here as the developed metrics identify molecules that bind to the minor groove, as this is the prefereable site of interaction on nucleic acids for most small molecules. Methylene blue is also an intercalator, but is unique in that at different ionic conditions, it may also interact with the major groove of AT rich DNA [112, 122]. Methyl green has been shown to prefer AT rich sequences and bind to the major groove of many different sequences of DNA[122, 152]. The compound pjp407 is a 2-phenylnapthalene derivative that has structure supporting intercalation [153]. Compound pjp72 appears similar in structure to some of the amidofluorenones that were previously discussed which suggests it possesses an intercalation or threading intercalation binding mechanism. Pm008 has a structure suggesting either groove binding or intercalation [121]. Quinacrine has long been utilized as an anti-malarial drug and is thought to act predominantly by intercalation. Sampangine is another anti-malarial and anti-fungal drug and the fused ring structure suggests it binds by classical intercalation [154-155].

The Surflex metrics appear to classify many of the compounds as exclusively intercalators, as there are many positive scores observed for the intercalation sites with few positive scores for the groove binding sites (Figures 58 and 59). Interestingly,

Figure 58. Surflex-Dock (top) and Autodock (bottom) docking scores for the Other Compound set from the 67 Compound Set, after application of the groove binding metrics. The results shown are for the groove binding sites in the nucleic acid library.

Figure 58. Surflex-Dock (top) and Autodock (bottom) docking scores for the Other Compound set from the 67 Compound Set, <u>after</u> application of the <u>groove binding</u> metrics. The results shown are for the <u>groove binding</u> sites in the nucleic acid library.

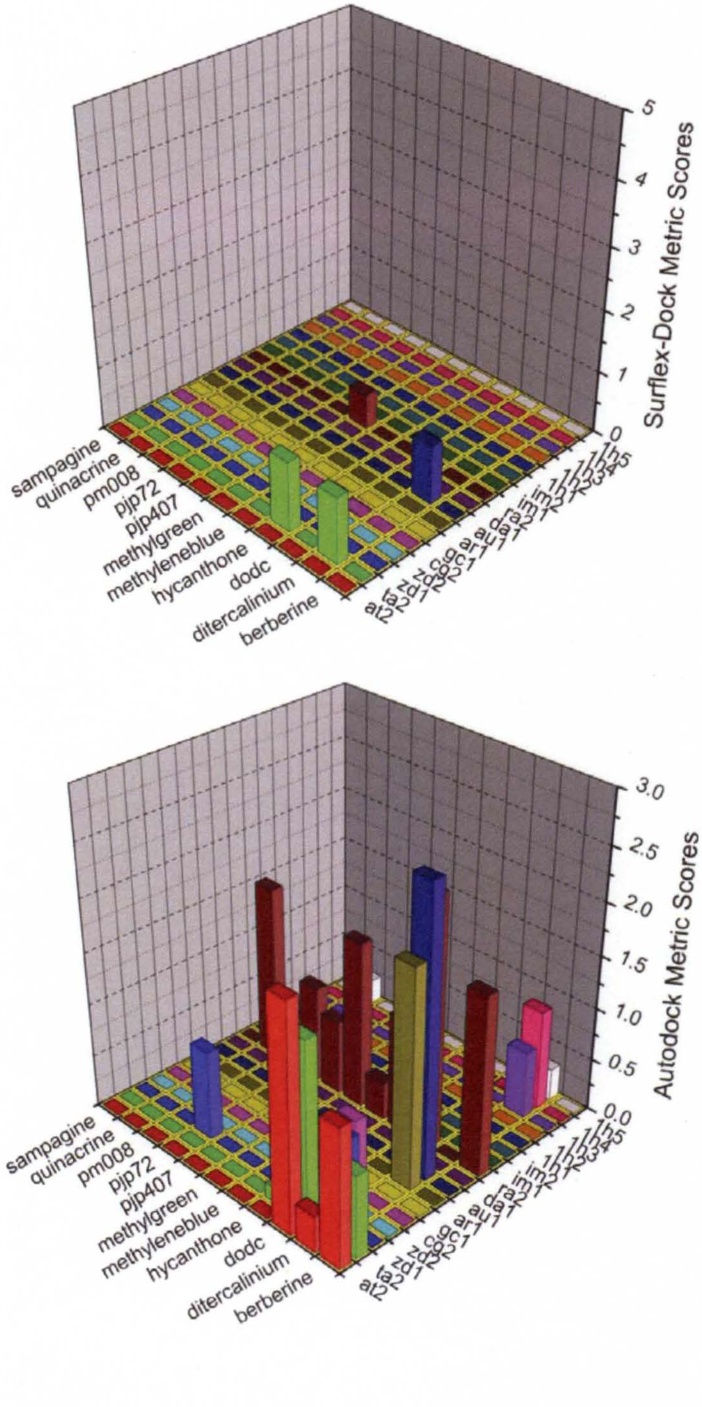
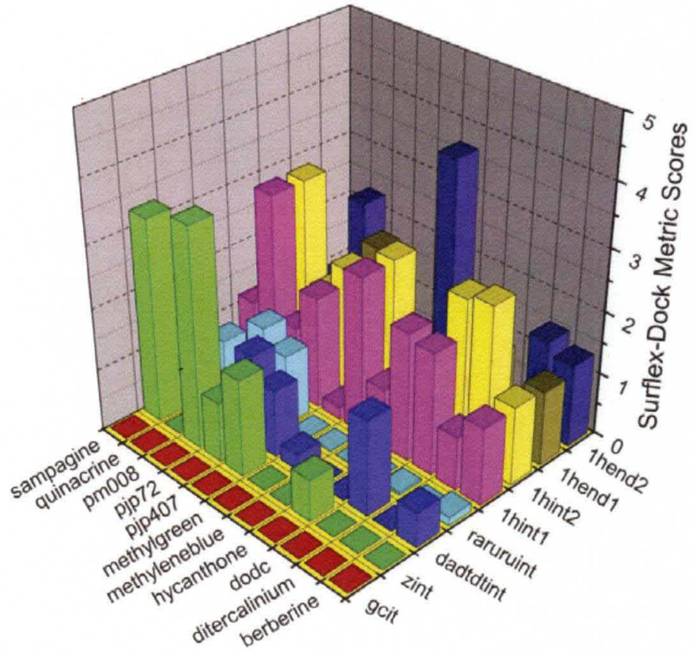
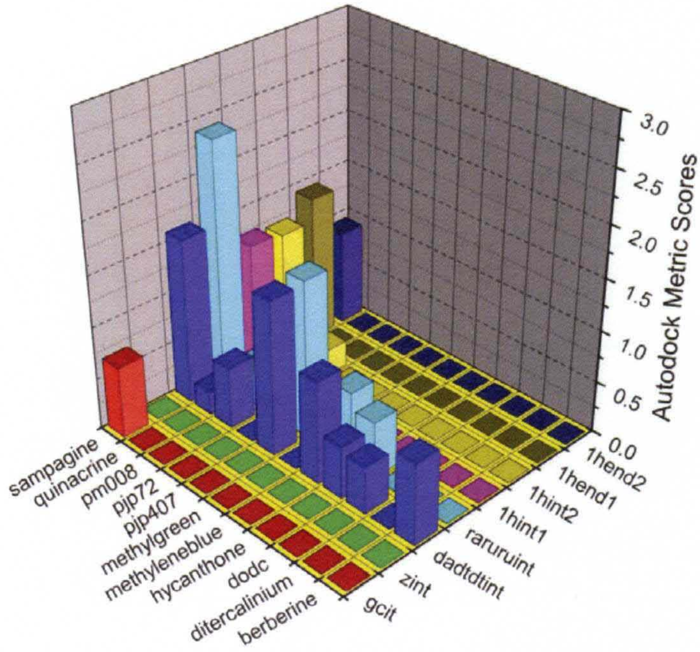


Figure 59. Surflex-Dock (top) and Autodock (bottom) docking scores for the Other Compound Set from the 67 Compound Set, after application of the intercalation metrics. The results shown are for the intercalation sites in the nucleic acid library.

Figure 59. Surflex-Dock (top) and Autodock (bottom) docking scores for the Other Compound Set from the 67 Compound Set, <u>after</u> application of the <u>intercalation</u> metrics. The results shown are for the <u>intercalation</u> sites in the nucleic acid library.





ditercalinium and methylgreen are predicted to have some positive groove scores, suggesting groove binding is also involved in their mode of binding which is consistent with known data. While Surflex successfully predicts many of these compounds to be intercalators, it does a marginal job of predicting structural preference, as many of the compounds bind triplex DNA, while duplex and quadruplex intercalation sites are the preferred *in silico* binding sites. Autodock shows many of the compounds having some groove binding character to different nucleic acids, which is not surprising given the structural diversity. However, almost all of the compounds also are predicted to intercalate into triplex nucleic acids (Figure 59). This finding is significant, as it is most consistent with the mechanism of action of many of these ligands as well as the structural specificity. In total, these results suggest that even for a heterogeneous group of ligands such as those shown here, Surflex and Autodock can generally successfully predict the binding mechanism and sometimes predict structural preference of the ligands.

Comparison and Limitations of Surflex and Autodock Performance and Metrics. The data presented here allow some comparisons to be made between Surflex and Autodock about the performance of these docking programs as well as limitations of the software. For both programs, the metrics that were developed were able to generally differentiate groove binding small molecules (Pentamidine and Distamycin) from intercalators (Daunorubicin and Ellipticine) in the Positive Control set. A similar finding was observed when performing robustness testing in the Validation and 67 Compound library sets (Table 8). Moreover, in some cases, even sequence and structural selectivity were successfully predicted *in silico* and generally mimic the actual *in vitro* competition dialysis data. Perhaps the best example of the success of the metrics are the triplex

Table 8. Comparison of Software for the Prediction of Compound Binding Mechanism After Application of the *In Silico* Metrics

Table 8. Comparison of Software for Predicting Binding Mechanism After Application of the *In Silico* Metrics

Compound Classification	**Number of Compounds**	**Primary Binding Mechanism**[1]	**Surflex Results**[2]	**Autodock Results**[2]
Ethidium Bromide Derivatives	9	Intercalation	🟨 4 of 9 correct	🟨 5 of 9 correct
Acridine Derivatives	6	Intercalation	🟩 5 out of 6 correct	🟨 3 of 6 correct
Aromatic Diamidine Derivatives	3	Intercalation	🟨 1 out of 3 correct	🟩 2 out of 3 correct
Cyclic Aromatic Derivatives	6	Intercalation	🟩 6 out of 6 correct	🟨 2 of 6 correct
Bis-quinoline Derivatives	8	Intercalation & Groove Binding	🟩 6 out of 8 correct	🟥 0 out of 8 correct
Amidoanthraquinones (non-aromatic side chains)	4	Intercalation & Groove Binding	🟩 4 out of 4 correct	🟥 1 out of 4 correct
Naphthoflavones	2	Intercalation	🟩 2 out of 2 correct	🟩 2 out of 2 correct
Amidofluorenone Derivatives	4	Intercalation & Groove Binding	🟩 4 out of 4 correct	🟥 0 out of 4 correct
Other Compounds	11	Intercalation & Groove Binding	🟩 9 out of 11 correct	🟨 5 out of 11 correct

1: Binding mechanism was determined from the cited literature publications.

2: The color scheme represents the accuracy to which the binding mechanisms are predicted: Good (Green, ≥66.6%), Average (Yellow, 33.3-66.5%) or Poor (Red, ≤33.2%). For compounds that bind exclusively by Intercalation or Groove Binding, the compounds were classified as predicted correctly if only positive scores are observed for the primary mechanism of binding and no other positive scores are observed for another mechanism of binding. In the case of compounds where both Intercalation and Groove Binding is significant, the compounds were classified as predicted correctly if the compounds showed positive scores for both intercalation and groove binding sites.

selective naphthoflavones, whose behavior was generally accurately predicted by the *in silico* metrics. We generally found Surflex to outperform Autodock with respect to predicting the binding mechanism while both programs had modest success at predicting sequence selectivity of the compounds. The success with Surflex was somewhat surprising, as Surflex is not parameterized for DNA and operates largely by-shaped based complementarity. On the other hand, Autodock has been used previously for targeting DNA and is specifically parameterized for nucleic acids [34]. This may explain the ability of Autodock in some cases to outperform Surflex when predicting sequence selectivity of some classes of compounds in the 67 compound library such as the Naphthoflavones. Our general findings, however, support the use of Surflex for further study as a molecular docking tool to use to target nucleic acid for small molecule discovery.

Based on the results from this study, there were several limitations of these software that require further elaboration. In general, the docking programs appeared to have a more difficult time predicting sequence and structural selectivity rather than predicting the binding mechanism, based on the *in silico* metrics developed here. We suspect that this is likely because even small structural changes can dramatically impact sequence and structural selectivity of a small molecule for nucleic acids. The docking programs also appear to have the most difficult time predicting the binding mechanism of larger molecules that bind by both intercalation and groove binding mechanisms. An example are the bis-quinolines which possess aromatic core scaffolds that can intercalate as well as a linker chain that bind into the grooves of the nucleic acids. Additionally, some molecules such as methylene blue bind to atypical sites such as the major groove of duplex DNA instead of the minor groove. The metrics as developed here may be inappropriate for classification of these few compounds as the metrics focus on predicting molecules that bind to the minor groove which is typically where small molecules interact with DNA. Finally, we emphasize here that the array of nucleic acids is by no means all encompassing but was merely used as this is the same array used for the competition dialysis assay and facilitated the comparison of the 67 compound library *in silico* and *in vitro* results. The array as constructed here is just a starting point. In the future, structures can be added to the *in silico* array as they become available that will hopefully add more diversity and power to this *in silico* approach.

Conclusions

Predicting how small molecules can interact with nucleic acids is crucial to discovering new compounds that target biologically relevant nucleic acids. Bourdouxhe-Housiaux *et al.* outline three criteria that can profoundly impact the biological activity of compounds that interact with DNA; (1) mechanism of ligand interaction with nucleic acids; (2) sequence specificity and (3) kinetics of association and dissociation [106]. We address points (1) and (2) here by inventing a novel approach to predict *in silico* how small molecules interact with nucleic acids. *In silico* rules were developed based on the docking of four small molecules (Daunorubicin, Distamycin, Ellipticine and Pentamidine) to an array of nucleic acids that allowed for the classification of these small molecules as either intercalators or groove binders. These metrics were tested for robustness on several compounds that our lab has discovered as well as a 67 compound library, for which extensive competition dialysis exists. The results of the testing with the 67 compound library confirmed that the Surflex and Autodock metrics are generally more successful at predicting binding mechanisms rather than sequence selectivity. A logical extension of this work is to utilize the metrics for the discovery of novel compounds that bind by a known mechanism to a specific nucleic acid. We envision that this would be accomplished by large scale virtual screening of millions of compounds to a nucleic acid target of interest. Top hits could then be screened against the *in silico* array of nucleic acids to check for unwanted binding and the metrics could be applied to elucidate the binding mechanism. As Surflex is approximately 10 times faster than Autodock under the conditions tested here, we describe in the next chapter of this work the use of Surflex with integrated selectivity metrics for the purposes of discovering new small molecules that bind nucleic acids.

CHAPTER IV

DISCOVERY OF NOVEL TRIPLE HELICAL DNA INTERCALATORS BY AN INTEGRATED VIRTUAL AND ACTUAL SCREENING PLATFORM

This chapter describes the use of ligand-based and structure-based virtual screening approaches for the purposes of discovering new small molecules that can specifically bind to triplex DNA. Our previous results in Chapters II and III supported the use of the virtual screening software Surflex for predicting the mechanism of action of small molecules that interact with nucleic acids. In this chapter, we use metrics to virtual screen millions of compounds to discover small molecules that specifically interact with triplex DNA by the mechanism of intercalation. My contribution to this work was the virtual screening experimental design and execution. Patricia Ragazzon, Ph.D. was responsible for the biophysical characterization experiments. The results of this work were published by P.A. Holt *et al* [64].

Virtual screening is an increasingly attractive way to discover new small molecules with potential medicinal value. We introduce a novel strategy that integrates use of the molecular docking software Surflex with experimental validation by the method of competition dialysis. This integrated approach was used to identify ligands that selectively bind to the triplex DNA poly(dA)-[poly(dT)]$_2$. A library containing approximately 2 million ligands was screened to identify compounds with chemical and structural similarity to a known triplex intercalator, the napthylquinoline MHQ-12. Further molecular docking studies using compounds with high structural similarity resulted in the discovery of two compounds that were then demonstrated by competition dialysis to have a superior affinity and selectivity for the triplex nucleic acid than MHQ-12. One of the compounds has a different chemical backbone than MHQ-12, which demonstrates the ability of this strategy to "scaffold hop" and to identify small molecules with novel binding properties. Biophysical characterization of these compounds by circular dichroism and thermal denaturation studies confirmed their binding mode and selectivity. These studies provide a proof-of-principle for our integrated screening strategy, and suggest that this platform may be extended to discover new compounds that target therapeutically and physiologically relevant nucleic acid morphologies.

DISCOVERY OF NOVEL TRIPLE HELICAL DNA INTERCALATORS BY AN INTEGRATED VIRTUAL AND ACTUAL SCREENING PLATFORM

Patrick A. Holt, Patricia Ragazzon, Lucjan Strekowski, Jonathan B. Chaires and John O. Trent. Discovery of Novel Triplex Helical DNA Intercalators by an Integrated Virtual and Actual Screening Platform. Nucleic Acids Res. 2009 Mar; **37** (4):1280-7. Epub 2009 Jan 9.

Introduction

Triple helical nucleic acid forming sequences have become a source of increasing interest as a way to interfere with DNA transcription and modulate gene expression [56-57, 156]. Several approaches attempt to use triplex nucleic acids to interfere with the transcription of genes, through either inducing the formation of triplex or stabilizing existing triplex nucleic acids. The former approach is the so-called "anti-gene" approach and involves the administration of triplex forming oligonucleotides (TFOs), which are short sequences of nucleic acids that can bind to the major groove of duplex nucleic acids in genes and promote the formation of triplex structures [58, 157]. TFOs have already been successful in reducing transcription of the *c-myc* oncogene that is located in the promoter site of genes [58, 158]. However, there are currently significant challenges associated with the use of TFOs and triplex structures in general. First, TFOs have significantly lower activity in cell-based systems, compared to *in vitro* systems [159]. This has been ascribed to many factors including improper cellular localization, degradation of the oligonucleotide, or lack of accessilibity to chromatin wrapped DNA [159-160]. A second limitation is the inherent low stability of many triplex structures [161-162]. The latter limitation is the focus of this work where we demonstrate the use of a novel virtual and actual screening platform for identifying several compounds that can selectively bind to and stabilize a triplex nucleic acid structure. These newly identified small molecules could be used to target triplex structures in several ways. First, the small molecule could stabilize pre-existing triplex structures *in vivo*. The small molecules could be used in an adjuvant setting with TFOs to increase stability, or alternatively the small molecules can be linked to TFOs to enhance the stability of newly formed triplex structures [160]. Either of these approaches could be used to control gene expression. These capabilities make these small molecules potentially clinically relevant for treating cancer and other diseases that are closely linked with abnormal gene expression.

Several small molecules are known to intercalate into and stabilize triplex nucleic acids including coralyne, benzo[*e*]pyridoindoles (B*e*PI), benzo[*g*]pyridoindoles (B*g*PI), dibenzophenanthrolines, and anthraquinones [136, 162-165]. One of the most selective and extensively studied classes are the napthylquinolines, which have been shown to intercalate into the TAT DNA triplex, poly(dA)-$[\text{poly(dT)}]_2$ [166-168]. Chaires *et al.* performed an extensive study that characterized the selectivity and affinity of 14 napthylquinoline derivatives [44]. The ligand MHQ-12 emerged from that study as the compound with the highest affinity and greatest selectivity for the poly(dA)-$[\text{poly(dT)}]_2$ triplex. While this approach for the discovery of triplex-selective ligands was successful, it is a laborious and time-consuming process. We propose a novel alternative approach for finding ligands that target a particular structure in which virtual screening is used to identify promising ligand candidates followed by validation using competition dialysis. We demonstrate here that this approach can identify small molecules that intercalate into poly(dA)-[poly(dT)]$_2$ with higher selectivity and affinity than MHQ-12. A significant result of this approach is that a small molecule with a substantially different molecular scaffold was identified that has superior affinity and selectivity for triplex DNA compared to MHQ-12. This strategy thus provides a new platform for identifying promising small molecule drugs against nucleic acid targets.

Virtual screening using molecular similarity and docking methods is becoming an increasingly important and economical approach to identify small molecules drug candidates [24]. While there are numerous studies using such screening methods for targeting proteins, far fewer virtual screening efforts have been performed targeting nucleic acids. The few studies that have been performed targeting nucleic acids have produced promising results. They have shown that screening methods can accurately reproduce crystallographic structures of ligand-nucleic acid complexes using a variety of docking programs including DOCK, Autodock and Surflex [33-34, 63]. Our virtual screening approach uses both ligand and structure-based discovery principles to select ligands from a commercially available library that bind to poly(dA)-[poly(dT)]$_2$ with higher affinity and selectivity than MHQ-12. Initial virtual screening is performed with Surflex-Sim, which is a ligand similarity-based software program that has superior performance compared to most traditional 2D similarity methods [169]. This program is an effective tool to rapidly prescreen large virtual compound libraries to enrich for structurally similar ligands [169]. Surflex-Sim maximizes 3D morphological similarity and alignment of a test ligand to the control ligand, which in this work was MHQ-12 [169-171]. The quantitative metric that is used for evaluating Surflex-Sim results is the Surflex-Sim score, which embodies an all atom comparison and alignment of the test ligand with the control ligand. The top ranked Surflex-Sim results were used for structure-based docking studies to dock the ligand to the intercalation site and the three grooves (major-major, major-minor and minor) of the triplex structure using the docking program Surflex-Dock [44]. Surflex-Dock performs docking of test ligands to a "protomol" or idealized representation of the binding site on the nucleic acid target. The ligands are docked to the target and the poses are ranked by a Surflex Raw Score (SRS) which consists of an affinity score of the ligand for the target [171]. This sequential combination of Surflex-Sim followed by Surflex-Dock produced several ligands that had hypothesized higher binding affinity and selectivity for the triplex intercalation site, compared to MHQ-12.

A critically important step after virtual screening is validation by experimental testing of the top candidates. To accomplish this, competition dialysis was employed because of its extensive use to determine the selectivity and affinity of a small molecule for single stranded, duplex, triplex and quadruplex nucleic acid targets [7, 10, 130, 164, 172-179]. The advantage of competition dialysis is that it is not limited to the target sequence, or a simple comparison with one other form of DNA, but with as many nucleic acid forms as are included in the assay. Competition dialysis involves dialyzing solutions of an array of nucleic acid sequences and structures against a common solution containing a test ligand [10]. The solution is allowed to reach equilibrium, and the amount of ligand that is bound to each nucleic acid is measured using either fluorescence or absorbance [10]. Comparison of the total and relative amounts of ligand bound to each nucleic acid assesses the affinity and selectivity, respectively, of the ligand for any included nucleic acid. Competition dialysis testing is used here to validate the affinity and selectivity of the top virtual screening hits. Circular dichroism and thermal denaturation were used for further characterization of the triplex binding of the top virtual screening candidate hits.

By using this integrated approach we have identified small molecules that have higher selectivity and affinity for triplex poly(dA)-[poly(dT)]$_2$ than the original molecule, MHQ-12, and which are among the most selective and tightest triplex binding molecules reported to date.

Materials and Methods

Virtual Library Construction. The triplex-selective ligand MHQ-12 was constructed and hydrogen atoms were added using Macromodel (version 7.0). The ligand was energetically minimized using a sequential combination of 2000 iterations of a Steepest Descent algorithm followed by 2000 iterations of a Polak-Ribier Conjugate Gradient (PRCG) algorithm. AMBER ligand atom types were assigned using Sybyl (version 7.3). The program Antechamber in the software suite Amber (version 8) was used to calculate AM1-BCC charges for the ligand and to convert to a MOL2 file format. A virtual set of 1.962 million ZINC compounds in MOL2 format were obtained from the ZINC 2007 "all-purchasable" subset of ligands from the University of California San Francisco (UCSF) [180]. The triplex nucleic acid structure poly(dA)-[poly(dT)]$_2$ with an intercalation site was constructed using B-type parallel triplex with X-ray structural intercalation site backbone fragments [Protein Data Bank entry 1p20.ent] and minimized holding the heavy atoms fixed [181].

Surflex Methods. The program Surflex (version 2.11) containing both the Surflex-Sim molecular similarity and the Surflex-Dock molecular docking programs was run on 30 AMD Opteron 246 processors (2.0 GHz) running the Linux Red Hat operating system for all virtual screening experiments. Surflex-Sim experiments were performed using the "align_list" function to compare the MHQ-12 triplex selective ligand against 1.962 million compounds in the ZINC library. The top 350 ligands, ranked according to the highest Surflex-Sim score, were selected as candidates for Surflex-Dock studies and were extracted as individual MOL2 files from the library files using in house PERL scripts. For the Surflex-Dock experiments, four protomols were generated to cover the major-major groove, major-minor groove, minor groove and intercalation sites of the triplex nucleic acid, using the same methods previously described [63]. The "proto_bloat" function was set to accommodate all reasonable interactions of the protomols with the triplex target sites. Docking of the ligands to the target was performed using a whole molecule approach, as described previously [60, 63, 171]. The Surflex-Dock experiments involved docking each of the ligands to all four protomols individually, in separate experiments. Surflex-Dock was operated with parameters "Multistart 5" and "Random 5," which we have previously shown returned accurate top ranked docked poses for a set of small molecules to their respective nucleic acid targets [63]. The Surflex-Dock poses were ranked according to the highest Surflex Raw Score [97]. Surflex-Sim and Surflex-Dock poses were visualized using AutoDockTools (version

1.4.6). The properties of compounds 1 and 2 used in the QSAR analysis were generated with QikProp [105].

Compounds for Biophysical Testing. The highest ranked candidates identified by virtual screening were the ligands with ZINC identification numbers 632255 and 4623551, which will be referred to hereafter as compound 1 and compound 2, respectively. Compound 1, is 4-(4-methylpiperazino-2-(2-naphthyl)quinoline and was obtained from Sigma-Aldrich (Milwaukee, WI). Compound 2 is 1-phenyl-4-pyrrolidino-2,3-dihydro-1/H/-pyrrolo[2,3-/b/]quinoline and was obtained from Chemical Block (N.D. Zelinsky Institute of Organic Chemistry, Moscow, Russia). As positive controls, the known triplex selective ligands MHQ-15 and OZ-85H were synthesized as previously described [44].

Competition Dialysis Method. Competition dialysis experiments were done as previously described [7, 44, 173, 182]. The array of oligonucleotides used is given in Table 9. The buffer used consisted of Na_2HPO_4 (6 mM), NaH_2PO_4 (2 mM), NaCl (185mM), EDTA (0.1 mM), pH 7. All nucleic acid samples were of identical concentration of 75 μM, expressed in terms of monomeric unit (base pairs for duplex DNA, triplets for triplex DNA and tetrads for quadruplex DNA). At the end of the dialysis equilibration period, ligand concentrations were determined by fluorescence. A volume of 180 μl of each sample was carefully transferred into one well of a 96-well microtiter plate (Costar® cat# 3915; Corning Inc., Corning, NY). To each sample, 20 μl of a 10% (w/v) sodium dodecyl sulfate (SDS) stock solution was added to give a final concentration of 1% (w/v) SDS, sufficient to dissociate the ligand from the DNA structure and ensure that there are no complexities arising from differences in the optical properties of free and bound ligands. The total ligand concentration (C_t) within each dialysis well was determined

Table 9. Oligonucleotide Array for Competition Dialysis.

Table 9. Oligonucleotide Array for Competition Dialysis

Conformation	Nucleic acid	Nomenclature	λ(nm)	ε ($M^{-1}cm^{-1}$)	Monomeric unit
Single-stranded DNA	poly (dA)	dA	257	8600	Nucleotide
Single-stranded DNA	poly (dT)	dT	264	8520	Nucleotide
Double-stranded natural DNA	Clostridium Perfringens (GC 31%)	C. perf.	260	12476	base pair
Double-stranded natural DNA	Calf thymus (GC 42%)	C.T.	260	12824	base pair
Double-stranded natural DNA	Micrococcus lysodeiktus (GC 72%)	M. lys.	260	13846	base pair
Double-stranded DNA	Z-DNA	Z-DNA	254	16060	base pair
Double-stranded DNA	poly (dAdT)	dAdT	260	12000	base pair
Double-stranded DNA	poly (dAdT)-(dAdT)	dAT	262	13200	base pair
Double-stranded DNA	poly (dGdC)-(dGdC)	dGC	254	16800	base pair
Double-stranded RNA	poly (rArU)	rArU	260	14280	base pair
DNA-RNA hybrid	poly (rAdT)	rAdT	260	12460	base pair
Triplex DNA or RNA	poly(dA)-[poly(dT)]$_2$	dAdTdT	260	17200	Triplet
Quadruplex DNA	TG_4T	TG_4T	260	12672	Quartet

using a fluorescence standard curve for each tested ligand. Appropriate corrections were made for the small dilution resulting from the addition of the SDS stock solution. The free ligand concentration (C_f) was determined from an aliquot of the dialysate solution, which typically did not vary significantly from the initial 1 μM concentration. Fluorescence measurements were made using a Safire microplate reader (Tecan US, Durham, NC), with the following parameters: excitation and emission bandwidth, 10 nm, gain: 100. Compound OZ-85H: excitation/emission 260/494 nm, compound MHQ-15: excitation/emission 260/437 nm, compound 1 excitation/emission 260/490 nm, compound 2: excitation/emission 348/446 nm. The bound ligand concentration (C_b) was then determined by:

$$C_b = C_t - C_f \tag{7}$$

Binding constants, specificity sums (SS) and the ratio C_b/SS were calculated as follows [183]. Apparent binding constants for each structure or sequence, K_{app}, may be calculated by:

$$K_{app} = \frac{C_b}{(C_f)(S_{total} - C_b)} \tag{8}$$

where C_b is the amount of ligand bound, C_f is the free ligand concentration and S_{total} is the total nucleic acid concentration. By virtue of the experimental design used in the competition dialysis experiment, C_f = 1 μM and S_{total} = 75 μM, expressed in terms of the monomeric unit of the nucleic acid. The specificity sum, SS, is the sum of the normalized amounts bound to each nucleic acid species, *i*:

$$SS = \sum_i \frac{C_{b,i}}{C_{\max}} \tag{9}$$

where $C_{b,i}$ is the amount bound and C_{max} is the maximum amount bound to any species. The index *i* ranges from 1 to 13 in the current assay, corresponding to the 13 different nucleic acids. A SS value of 1 indicates absolute selectivity for one structure whereas a value of 13 indicates lack of selectivity. Information about compound binding affinity is a function of C_{max}. Thus, the ratio C_{max}/SS represents affinity and selectivity. If SS = 1, the maximal value of C_{max}/SS will be obtained and if SS = 13, the minimal value is the result. A high value of the C_{max}/SS ratio is representative of compounds with high binding affinity and selectivity.

CD Titration and Thermal Denaturation Methods. CD titrations were done as previously described, using a Jasco 810 spectropolarimeter. Instrument settings were: wavelength range (220-500 nm), scan rate (100 nm min^{-1}), averaging time (0.125 s), bandwidth (1 nm), number of scans (2), temperature (20 °C) [184]. The effect of ligands on the thermal denaturation of triplex DNA was studied using the exact protocol described previously [183].

Results and Discussion

Virtual Screening. The initial step in virtual screening was performing Surflex-Sim to determine which of the ligands in the library were most structurally similar to the known, triplex selective intercalator MHQ-12. Of the approximately 2 million ligands screened for similarity against MHQ-12, 350 ligands had a Surflex-Sim score of greater than 0.70 (range: 0.875 - 0.704) and were selected for Surflex-Docking studies. A cutoff Surflex-Sim score of 0.70 was selected based on previous studies which suggested that this is the lowest score where the ligand structure-function relationship is typically maintained [97].

The next step in the virtual screening process involved performing Surflex-Dock studies with the top 350 ranked Surflex-Sim ligands using the intercalation site and the three grooves (major-major, major-minor and minor) of the triplex as individual docking targets. Interestingly, MHQ-12 has the top Surflex Raw Score out of all 350 ligands that were docked to the intercalation site, which directly supports the ability of Surflex-Dock to successfully dock and rank a known selective triplex intercalator. We propose a new metric to evaluate the Surflex-Dock results, the "Normalized Surflex Raw Score (NSRS)". The rationale behind the normalization of the Surflex Raw Score is that the score for a ligand binding to a single site on a target measures only the interaction with that one site. However, a ligand may have multiple interaction sites on a particular target. Therefore for selectivity for a particular mode of binding, it is crucial to determine the binding of the ligand to the site of interest *relative* to the binding to other potential sites on the target. Since ligands interact with nucleic acids typically through either the groove-binding or intercalation, protomols were constructed at the three grooves and the intercalation site [75]. Binding of the ligand to the intercalation site *relative* to binding in the three grooves embodies the "normalized" affinity and specificity of the ligand for triplex intercalation. The following metric, which was first developed in Chapter III of this work, determines the NSRS for the intercalation site for each of the top 350 Surflex-Dock results:

$$\mathrm{NSRS}_{\text{intercalate site}} = \mathrm{SRS}_{\text{intercalate site}} - \text{Maximum } \mathrm{SRS}_{\text{major-major site, major-minor site, minor site}} \quad (10)$$

Ranking of the 350 intercalation site Surflex-Dock results by NSRS shows that only three ligands have a higher NSRS score than MHQ-12 (NSRS of 6.8) (Figure 60A). The ligands are LS-08 (Figure 60B), compound 1 (Figure 60E) and compound 2 (Figure 60F) and have NSRS values of 7.03, 7.34 and 7.39, respectively (Figure 60) [185]. Interestingly, LS-08 (Figure 60B) which was identified by our virtual screening methodology, was previously tested by Chaires *et al.* and shown to be highly triplex selective, which adds validity to our virtual screening approach used to identify triplex selective ligands [44, 183]. Based on the NSRS values, compounds 1 (Figure 60E) and 2 (Figure 60F) were hypothesized to have superior affinity and selectivity for binding to the triplex nucleic acid, and were tested by competition dialysis. Two known triplex selective compounds, MHQ-15 (Figure 60C) and OZ-85H (Figure 60D) served as positive controls, as these compounds have been extensively studied and characterized [44]. Biophysical characterization was performed by circular dichroism and thermal denaturation to assess the ability of the compounds to intercalate into the DNA triplex.

Competition Dialysis. The results of the competition dialysis experiments are shown in Figure 61. It is visually apparent that compounds 1 and 2 have a much higher affinity for the TAT triplex than the two positive control reference compounds, MHQ-15 and OZ-85H. The competition dialysis results for MHQ-12 have previously been described in detail, and this compound has a SS of 1.32 and a Cmax/SS ratio of 8.93 [44]. Determination of the SS (Table 10) for compounds 1 and 2 demonstrates superior triplex selectivity compared to OZ-85H but slightly less selectivity than MHQ-12 and MHQ-15. However, the significantly higher binding affinities of compounds 1 and 2 translates to

Figure 60. Chemical structures of the ligands used in virtual screening and competition dialysis Experiments. (A) MHQ-12, (B) LS-08, (C) MHQ-15, (D) OZ-85H, (E) compound 1 and (F) compound 2.

Figure 60. Chemical structures of the ligands used in virtual screening and competition dialysis experiments.

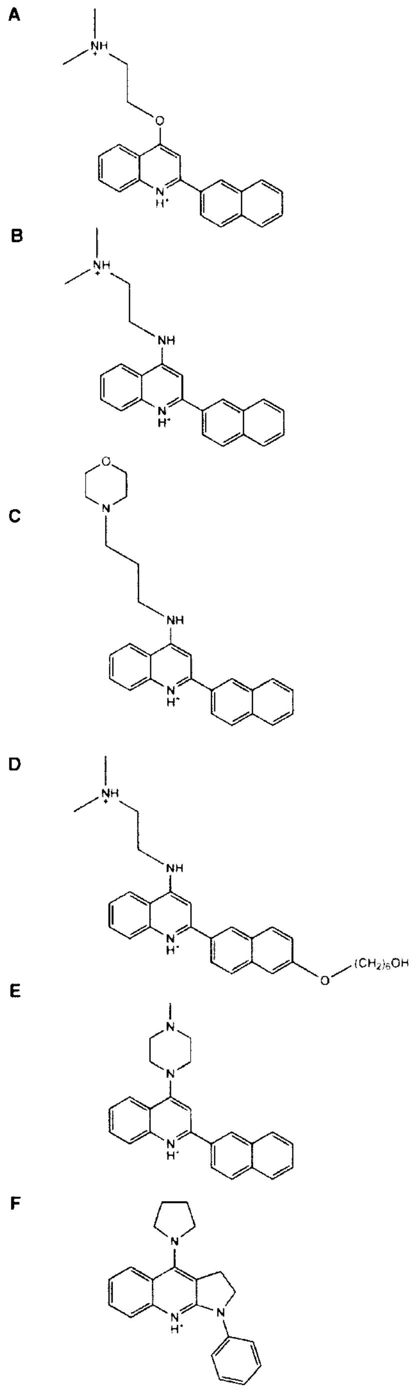


Figure 61. Competition dialysis results for MHQ-15, OZ-85H, compound 1 and compound 2. The concentration of bound ligand to each nucleic acid structure in the array is shown.

Figure 61. Competition dialysis results for MHQ-15, OZ-85H, compound 1 and compound 2.

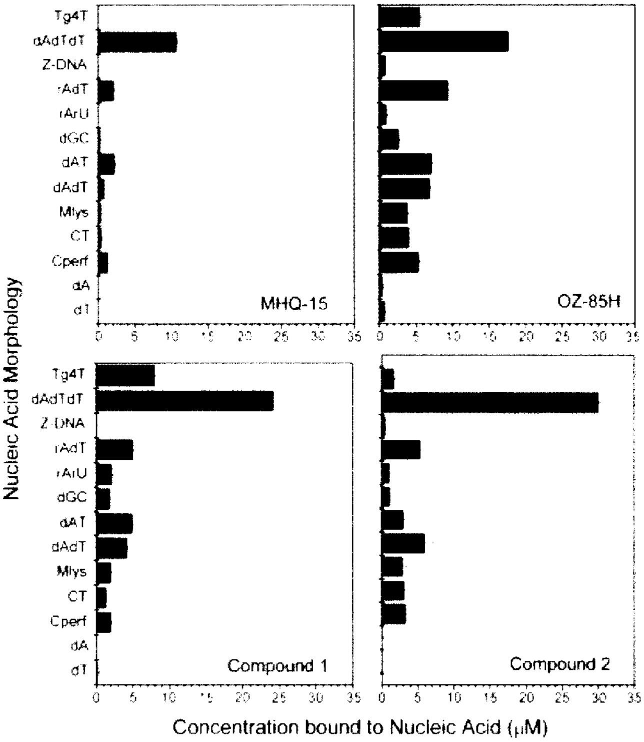


Table 10. Competition dialysis metric results for the positive controls, MHQ-15, OZ-85H, and the virtual screening top results, compounds 1 and 2.

Table 10. Competition dialysis metric results.

Test Ligand	**C_b (μM)**	**$K_{app}/10^5$ (M^{-1})**	**SS**	**C_{max} / SS (μM)**
MHQ-15	10.7	1.7	1.66	6.44
OZ-85H	17.6	3.1	3.69	4.77
Compound 1	24.2	4.8	2.30	10.47
Compound 2	30.0	6.7	1.92	15.63

much higher Cmax/SS values than MHQ-12, MHQ-15 or OZ-85H. The Cmax/SS ratio for compounds 1 and 2 is significant as it suggests that compounds 1 and 2 have a superior combination of binding affinity and selectivity compared to the reference compounds. These results validate the virtual screening approach, and show that the method can be used to identify compounds with high affinity and selectivity for a target nucleic acid, in this case the DNA TAT triplex.

Circular Dichroism. The interaction of compounds 1 and 2 with DNA was studied by circular dichroism (Figure 62). Both compounds show pronounced induced circular dichroism (ICD) in the presence of triplex DNA. The ICD is in a spectral range where the compounds absorb light but the DNA does not. This ICD is unambiguous proof of the ligand binding to triplex DNA. For both compounds 1 and 2, the ICD is negative in sign, and relatively weak in magnitude. Such behavior is consistent with an intercalative binding mode, although the mode of binding can only be definitively established by high-resolution experimental structural analysis [186].

Thermal denaturation studies. Figure 63 shows the effects of compounds 1 and 2 on the thermal denaturation of the TAT triplex. In the absence of added ligand, two transitions are seen, corresponding to the melting of the third strand (~ 30°C) and the duplex (~ 70°C). Titration with both ligands results in a clear elevation of the first transition, indicating stabilization of the triplex. The effect is maximal at saturating concentrations of ligand (1:1, ligand:triplet), where melting of the triplex coalesces with duplex melting. Melting of the triplex is stabilized by ~ 40°C indicating tight binding of both compounds. Neither compound 1 nor compound 2 alter the transition temperature of the duplex form

Figure 62. Induced Circular Dichroism results for (A) compound 1 and (B) compound 2. (A) Spectra are shown for a ligand concentration of 45 μM in the presence of triplex DNA ranging from 5 μM to 450 μM triplets. (B) Spectra are shown for a ligand concentration of 22.5 μM in the presence of triplex DNA ranging from 2.25 μM to 225 μM triplets.

Figure 62. Induced Circular Dichroism results.

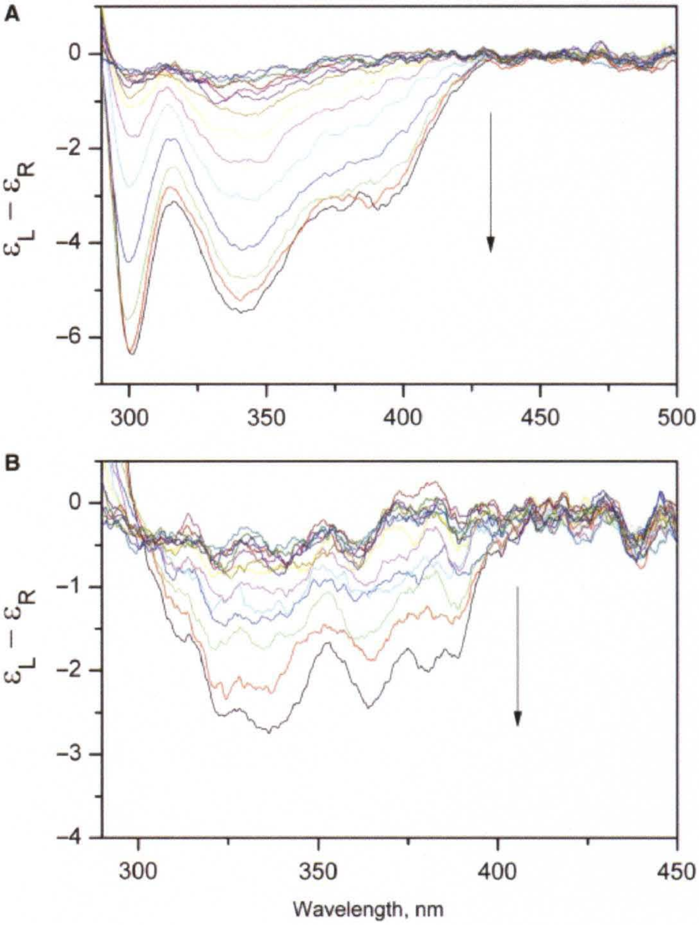
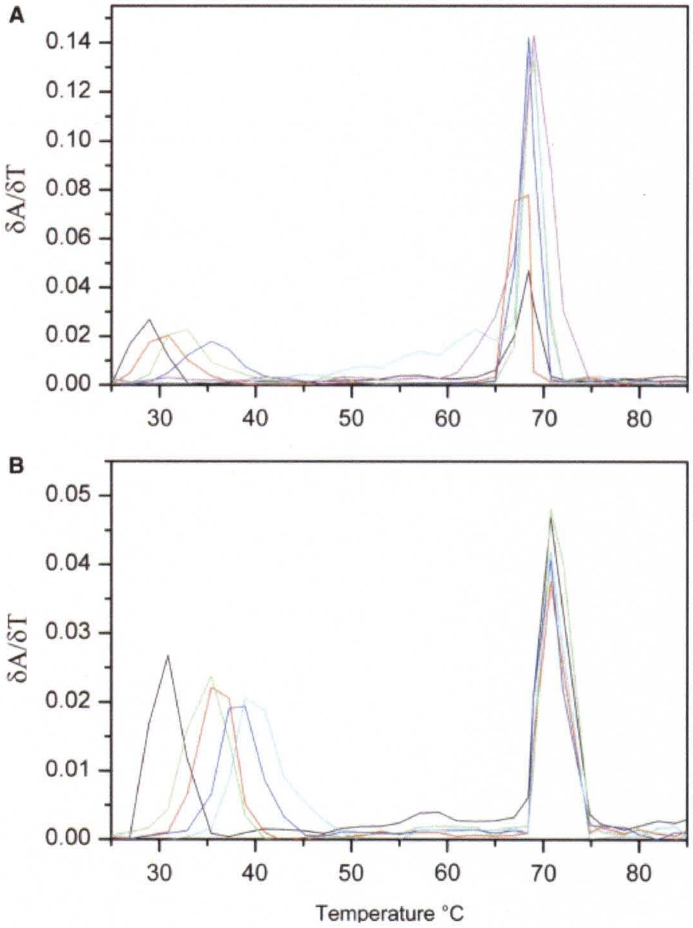


Figure 63. Thermal Melting results for (A) compound 1 and (B) compound 2. Derivative melting curves were obtained using 32 μM triplex DNA and ligand concentrations ranging from 0 - 16 μM (A) or 0 -32 μM (B). The peak near 30°C is for the melting of the third stand, while that near 70°C is for melting of the duplex.

Figure 63. Thermal Melting results.



to any appreciable extent, an observation that is fully consistent with the weak binding to duplex seen in competition dialysis experiments (Figure 61).

Validation of QSAR. In the previous study of naphthylquinoline binding to triplex DNA [44], a QSAR was derived from competition dialysis binding data. The best three-term QSAR to emerge was:

$$\begin{aligned}\log K_{app} &= 0.00264(\pm 0.00065)SASA - 0.693(\pm 0.125)EA - 0.196(\pm 0.02)HB_a \\ &+4.66(\pm 0.44) \\ N&=14; R=0.959; RMSE=0.130; F=49.84; P=0.0001\end{aligned} \tag{11}$$

In this relationship, log K_{app} is the logarithm of the apparent binding constant (Table 10), SASA is the total solvent accessible surface area in $Å^2$, EA is electron affinity in eV and HB_a is the number of hydrogen bond acceptors. The physical meaning of this is as follows. As SASA increases, log K_{app} increases in magnitude, indicating higher affinity for triplex DNA. Increases in the magnitudes of EA and HB_a result in decreasing binding affinity. Increasing the solvent accessible surface areas of naphthylquinoline compounds results in higher affinity for the triplex. Greater electron affinity and more hydrogen bond acceptors reduce the affinity of naphthylquinolines for triplex DNA.

Binding data obtained for compounds 1 and 2 in this study validate the published QSAR. The molecular descriptors SASA, EA and HB_a were calculated using QikProp, and substituted into equation (11). For compound 1, log K_a = 5.07 was predicted, compared to a measured value of log K_{obs} = 5.68. For compound 2, log K values of 5.18 and 5.82 were calculated and observed, respectively. The differences in calculated and observed values correspond to a factor of about 4 in binding constants, an acceptable agreement for predictions from a QSAR.

Conclusion

This work demonstrates a novel strategy for discovering small molecules that can selectively bind to the triplex nucleic acid, poly(dA)-[poly(dT)]$_2$. Through the combination of virtual screening by Surflex and experimental validation by competition dialysis, compounds 1 and 2 were discovered. These compounds have the highest overall affinities and selectivities reported for triplex binders as determined by competition dialysis. Further biophysical characterization by circular dichroism and thermal melting confirmed the mechanism of action of these new compounds and verified the predictive nature of the virtual screening methodologies. Several aspects of the virtual screening results are noteworthy. First, the combination of a ligand-based (Surflex-Sim) with a structure-based approach (Surflex-Dock) proved to be a powerful and highly computationally efficient way to identify triplex selective small molecules, as Surflex-Sim is two orders of magnitude faster than Surflex-Dock. Second, our development of the NSRS metric, which can predict a particular mode of binding of triplex selective ligands with both similar and different (scaffold hopping) chemical scaffolds. This is significant as it has the potential to identify new classes of small molecules that may have much higher affinity and selectivity for a given nucleic acid target. Future work will focus on extending this integrated virtual and actual screening platform to target other nucleic acid structures that may hold medicinal value and physiological relevance.

CHAPTER V

DISCOVERY OF A G-QUADRUPLEX NUCLEIC ACID BINDING SMALL MOLECULE BY *IN SILICO* SCREENING AND MOLECULAR MODELING

This chapter describes the utilization of the validated software Surflex and Autodock for the purpose of discovering novel small molecules that can bind to G-quadruplex DNA structures. The targeting of G-quadruplex nucleic acids by virtual screening approaches remains vastly underexplored despite the potential anti-neoplastic use of small molecules that bind specifically to G-quadruplexes. We report here the development of a novel, structure-based virtual screening approach that uses the molecular docking software tools Surflex and Autodock to screen over 6.6 million compounds for their binding to a specific site within the human telomeric G-quadruplex AGGG(TTAGGG)$_3$. A novel compound with a scaffold unlike any previously reported was discovered *in silico*. The compound was demonstrated by spectroscopic and fluorescent biophysical methods to interact with the G-quadruplex by the specific mechanism predicted by the *in silico* screen. Models of the newly discovered compound interacting at various end-pasting sites on the human telomeric quadruplex were constructed which provides insights to the important ligand-nucleic acid interactions necessary for targeting quadruplex structures. The virtual screening approach as presented here may be applied to any number of nucleic acid targets to discover new compounds that may have medicinal benefit.

DISCOVERY OF A G-QUADRUPLEX NUCLEIC ACID BINDING SMALL MOLECULE BY *IN SILICO* SCREENING AND MOLECULAR MODELING

Patrick A. Holt, Robert Buscaglia, Jonathan B. Chaires, John O. Trent

Introduction

Discovering small molecules that bind to nucleic acids using high throughput *in silico* virtual screening continues to be a largely untapped area of computational research and drug discovery. Indeed, nucleic acid focused therapeutics currently represent only a few percent of marketed drugs, with the vast majority focused on protein targets [1]. This initial neglect of nucleic acids as viable targets appears partially due to the failure to appreciate the structural diversity and functional significance of nucleic acids. With advances in the understanding of the diverse structures of nucleic acids, there is now an increasing list of nucleic acid targets with physiological and *in vivo* relevance [8]. Among the most attractive nucleic acid targets are the G-quadruplexes, which are found in the human telomeric region of chromosomes and consist of the motif $(TTAGGG)_n$. These G-quadruplex nucleic acid structures have a novel mechanism of potentially inhibiting cancer cells replication [39, 53, 187]. Over 85% of cancer cells overexpress telomerase which allows cancer cells to maintain the ends of human telomeres and is ultimately responsible for cancer cell immortality [54]. G-quadruplex structures have been shown to destabilize telomerase from the telomere, resulting in decreasing cancer cell life [55]. Thus, these quadruplexes have become a source of great interest for the identification of small molecules that may bind and stabilize the structures *in vivo*, and inhibit telomerase activity. Efforts to discover quadruplex binding small molecules have so far been modestly fruitful, and highlighted by the movement of Quarfloxin (owned by Cylene Pharmaceuticals) into humans in planned clinical trials.

In addition to G-quadruplex in the human telomeric region of chromosomes, G-quadruplexes also exist with increased frequency in the promoter regions of many genes. It appears that oncogene promoter regions contain potential quadruplex-forming sequences at a statistically significant increased rate, such as *c-myc*, *bcl-2* and *VEGF* [188-189]. The *c-myc* promoter in particular has gained attention as its overexpression is strongly associated with cancer development. There is increasing evidence suggesting that G-quadruplexes play a role in the regulation and modulation of oncogene transcription and the G-quadruplexes have increasingly been the focus of small molecule targeted approaches. In the case of *c-myc*, the small molecule $TmPyP_4$ has been shown to stabilize quadruplex structures located in the nuclease hypersensitivity element III_1 (NHE) area of the promoter, which controls >80% of c-myc gene transcription [188]. Ultimately, TmPyP4 is able to effectively inhibit gene transcription by stabilizing the quadruplex [108, 189]. This emphasizes that G-quadruplexes in promoters as well as in the human telomeric region are attractive structures for small molecule targeting.

In spite of the promise of small molecules that may bind to G-quadruplex targets, in particular the human telomeric G-quadruplexes, there are very few published reports of large scale *in silico* molecular docking approaches to discover new small molecules that can bind to these targets [34, 36-37]. The studies that have been performed appear to screen only limited numbers of compounds, typically on the order of thousands of compounds [190]. A significant number on the order of tens of millions of *in silico* small molecules that are currently available that may bind to these targets have yet to be explored. The lack of computational structure-based small molecule discovery in this research area exists for several reasons. First, while the human telomeric sequence AGGG(TTAGGG)$_3$ is of intense interest and has been studied in depth *in vitro*, there remains great controversy as to the actual structure that this sequence adopts under physiologically relevant solutions in K^+. A number of structures and sequence variants have been reported which emphasizes the unique polymorphism of the human telomeric and other closely related sequences [191]. Published X-ray crystallographic structures suggest this sequence forms an all-parallel "propeller" shape [191-193]. However, an increasing body of evidence suggests that in fact, this structure may not represent the "correct" structure(s) under physiological conditions in solution [194-195]. Unfortunately, an NMR solution structure of the 22mer human telomeric sequence has not been published although a number of variant sequences containing this human telomeric sequence suggest that the human telomere adopts a so-called Hybrid or parallel/anti-parallel structure under physiological relevant conditions [193-194, 196-197].

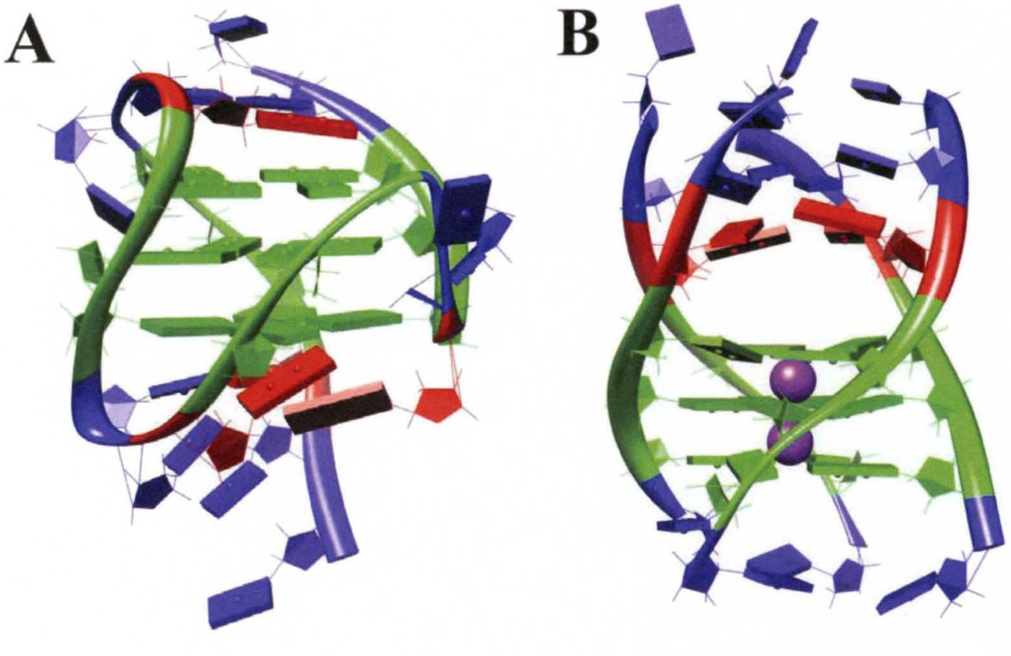
Another reason for the lack of *in silico* structure-based targeting of the human telomeric quadruplex is that molecular docking software in large part has been developed, parameterized and validated almost exclusively for protein targets, and may not appropriately consider the unique properties of nucleic acids. Also, previous computational studies have focused on rationalizing known data, which is an important but different type of experiment than using the software to discover new nucleic acid binding small molecules with novel scaffolds [34, 63]. Additionally, it remains to be seen whether large scale virtual screening of millions of small molecules to nucleic acid targets *in silico* is computationally feasible and whether the small molecules that are discovered possess the binding activity that was predicted *in vitro*. A rigorous *in vitro* validation is necessary to validate the predictive nature of computational approaches.

A final reason that *in silico* approaches have been ignored for targeting the human telomeric quadruplex appears to the focus of many research groups on "rational" drug discovery by derivatizing known quadruplex binding small molecules, such as TMPyP4 and other small molecules, to enhance binding to the human telomeric quadruplex structure [198-200]. Unfortunately, attempts to discover new small molecules with truly novel scaffolds that interact with G-quadruplexes by *in silico* based approaches have been severely limited. As a result, the ability of computational approaches to explore the full chemical space for new small molecule discovery remains underappreciated and underutilized. While these are only some of the challenges associated with targeting of quadruplex structures, two in particular are the main foci of this chapter. First, is how to select a relevant or representative human telomeric G-quadruplex structure for structure-based virtual screening. Second, is to determine if an integrated *in silico* molecular docking and *in vitro* testing platform can successfully discover and validate the binding of new small molecules to the human telomeric G-quadruplex structure.

An important issue that arises when targeting the human telomeric G-quadruplex is the choice of a representative structure to use for structure-based virtual screening. While there are known small molecules that appear to bind to the human telomeric sequence, the solved crystal structure to which these small molecules are bound may not be the "relevant" solution structure of the human telomeric quadruplex. For example, structures are available of small molecules bound to the "propeller" shaped all-parallel quadruplex even though the Hybrid type quadruplex structure is largely considered to be the "relevant" structure in solution [194-195, 201]. The lack of relevant solution structures with small molecules bound with this specific DNA structure has undoubtedly limited structure-based drug discovery approaches [202]. Small molecules are known to interact with the AGGG(TTAGGG)$_3$ quadruplex in three ways: first by groove interactions, second by intercalation between consecutive guanine tetrads and third by end-pasting, where the ligand is bounded on one side by the guanine quartet and on the other side by the loops of the quadruplex (Figure 64A) [203]. The end-pasting mechanism is of intense interest, as it is thought to confer selectivity for quadruplexes over other nucleic acid structures by taking into account both guanine quartet and loop interactions. Small molecules that interact in this manner are thought to stabilize G-quadruplexes, prevent replication by telomerase and result in decreased cancer cell proliferation [55]. We were interested in performing virtual screening experiments to discover small molecules that may end-paste on the AGGG(TTAGGG)$_3$ structure. However, a difficulty is identifying an *in silico* G-quadruplex structure in the RSCB PDB with a "representative" end-pasting site for targeting in which a small molecule is bound. For our purposes, it is preferable to use a virtual structure in which ligands are complexed as the ligands can be easily removed by computational methods and docking experiments performed without perturbing the nucleic acid structure. As we will show, the nucleic acid structure that was identified possesses an end-pasting site with strikingly

Figure 64. (A) A G-quadruplex structure (PDB ID 2JPZ) that contains the human telomeric repeat and shows potential ligand interaction sites in the grooves, intercalation sites and end-pasting sites. (B) The G-quadruplex (PDB ID 1NZM) with the sequence $(TTAGGT)_4$ with a Guanine-Adenine end-pasting site that was initially used for virtual screening with Autodock and Surflex. The RHPS4 ligand that is positioned in the end-pasting site is removed for clarity. Blue = Thymine, Red = Adenine, Green = Guanine and Purple = K^+ cations.

Figure 64. (A) A G-quadruplex structure (PDB ID 2JPZ) that contains the human telomeric repeat. (B) The G-quadruplex (PDB ID 1NZM) with the sequence $(TTAGGT)_4$ with a Guanine-Adenine end-pasting site that was initially used for virtual screening with Autodock and Surflex.



similar properties of the human telomeric end-pasting site and this end-pasting site served as the target for the virtual screening of millions of compounds.

After selection of a G-quadruplex nucleic acid structure for structure-based virtual screening, the next challenge is how to perform the molecular docking experiments. The use of molecular docking software to target nucleic acids has generally been limited almost entirely to proteins. However, our recent evidence suggests that two virtual screening software in particular, Surflex and Autodock, have great potential for molecular docking small molecules to nucleic acid targets [34, 63, 204]. We previously reported that both of these software accurately reproduced the crystal structures of a set of small molecules that interact with nucleic acids by both groove binding and intercalation [63]. However, the question remains whether the software can be used for virtual screening of millions of compounds to discover *new* small molecules that bind to a desirable target. Our previous results suggested that the accuracy of both software at reproducing known structures is comparable (under conditions previously tested), but Surflex is approximately 10 fold faster than Autodock and requires less file preparation for virtual screening [63]. Because of the complementary docking and scoring algorithms of Surflex and Autodock, however, we also investigate combining the power of both of these software into a single platform that is capable of novel small molecule discovery through a virtual screening strategy that is detailed below.

Even given these challenges, we report here the successful development of a high throughput *in silico* molecular docking platform that discovered a human telomeric quadruplex binding small molecule with a chemical scaffold unlike any reported to date in the literature. A quadruplex with the sequence $(TTAGGGT)_4$ that is complexed with a small molecule intercalated between a guanine-adenine tetrad was used as an *in silico* basis for the representative "end-pasting" site contained in the human telomeric quadruplex AGGG(TTAGGG)$_3$. A new virtual screening strategy is described that was successful and computationally plausible for screening millions of small molecules against a nucleic acid target. The top nine hits gleaned from the virtual screening studies were tested using spectroscopic and fluorescence based assays to validate the predicted *in silico* activity by *in vitro* testing. Finally, we computationally generated all possible end-pasting sites in two *in silico* RSCB PDB structures containing the human telomeric G-quadruplex AGGG(TTAGGG)$_3$ repeat and docked the newly discovered compound to the sites to assess nucleic-acid and small molecule interactions. The results show that the virtual screening platform, as described here, is predictive and capable of discovering new small molecules with a specific mechanism of interacting with G-quadruplex nucleic acids.

Experimental and Computational Methods

In silico Ligand Database and Nucleic Acid Target Preparation. A ZINC database consisting of approximately 6.6 million virtual small molecules was used for initial virtual screening studies against the nucleic acid (TTAGGGT)$_4$. These small molecules were the "Reference" subset of the 2008 "Drug-Like" dataset and are freely available for download from the University of California, San Fransisco [180]. The ligands have been named "Drug-Like" because of their adherence to Lipinski's rule of 5 to increase the chances that any hits will have higher oral bioavailability [31]. The small molecules were downloaded and used without any further modification from the initial procedures performed by UCSF which included protonation based on a pH 7 reference, 3D coordinate generation and partial charge assignment from AMSOL semi-empirical quantum calculations. The NMR determined G-quadruplex (TTAGGGT)$_4$ with PDB ID 1NZM was downloaded from the RSCB Protein Data Bank in PDB file format for use in the virtual screening *in silico* experiments. Sybyl v8.1 (Tripos, Inc.) was used to assign AMBER atom types and convert the file to MOL2 format in preparation for initial Surflex-Dock screening. The RHPS4 complexed ligand was removed prior to molecular docking experiments using Sybyl. All *in silico* virtual screening studies were performed on our server of 440 computers consisting of 2.66GHz Intel(R) Xeon(R) E5430 processors and required approximately 3 days to complete the Surflex-Dock experiments and 2 days to complete the Autodock experiments.

Surflex and Autodock Virtual Screening Methods. We have previously validated the use of Surflex-Dock and Autodock for targeting nucleic acids and these software are a logical choice for the discovery of new ligands against novel targets [63]. The end-pasting site on the nucleic acid (TTAGGGT)$_4$ was considered a model of the end-pasting site on the human telomeric quadruplex AGGG(TTAGGG)$_3$ and was targeted for molecular docking using both Surflex-Dock v2.2 (Tripos, Inc.) and Autodock v4.0 (Scripps). The end-pasting cavity was specified for Surflex-Dock v2.2 docking using a ligand-based approach. This involves using the existing RHPS4 ligand that was bound inside the (TTAGGGT)$_4$ end-pasting site to generate a Surflex-Dock "protomol" which guides the molecular docking of the *in silico* ligand library to the end-pasting site. The "protomol" was constructed by altering the "proto_bloat" and "proto_thresh" functions and visualized in Sybyl to ensure reasonable interactions in the end-pasting site. The significance of the protomol and the Surflex-Dock docking and scoring functions have been described in detail previously [60]. Briefly, the "protomol" consists of a series of small chemical fragments that model important forces in the nucleic acid pocket, including steric effects and hydrogen bond acceptor and donor groups. Each of the ligands in the *in silico* virtual library is fragmented, aligned against the protomol, and subsequently scored based on the interactions in the binding site. Surflex-Dock was performed using default options which in our previous experience is appropriate for rapidly screening databases of small molecules against nucleic acid targets.

The utilization of Autodock to target nucleic acid structures has also been previously described [34, 63-64, 204]. Autodock works by precomputing energy grids for a target [205]. A genetic algorithm such as the Lamarkian Genetic Algorithm is used to assess the interactions of the ligand with the pre-calculated energy grids until typically a specific number of energy evaluations is reached. The final top "pose" returned by Autodock is the computed lowest energy docked structure of the ligand with respect to the target. The highest scoring pose of the top 1% of ranked Surflex-Dock hits (approximately 66,000 small molecules) were extracted in MOL2 format and converted to PDBQT file format using the Python scripts included with Autodock. The G-quadruplex nucleic acid $(TTAGGGT)_4$ was prepared for Autodock by using AutoDockTools to convert the MOL2 to PDBQT file format. The extent of the Autodock grid maps was 66,64,40 points (X,Y,Z) with grid spacing distance of 0.375Å, and the grid centered on the end-pasting site. Autodock docking was performed by setting the number of docking runs to 5 and energy evaluations to 20 million energy evaluations, as we previously found these parameters to be optimal for docking of small molecules to nucleic acids [63]. After completion of Autodock docking, the Surflex and Autodock results were re-ranked using a Ranked Consensus Scoring (RCS) function as follows [206]:

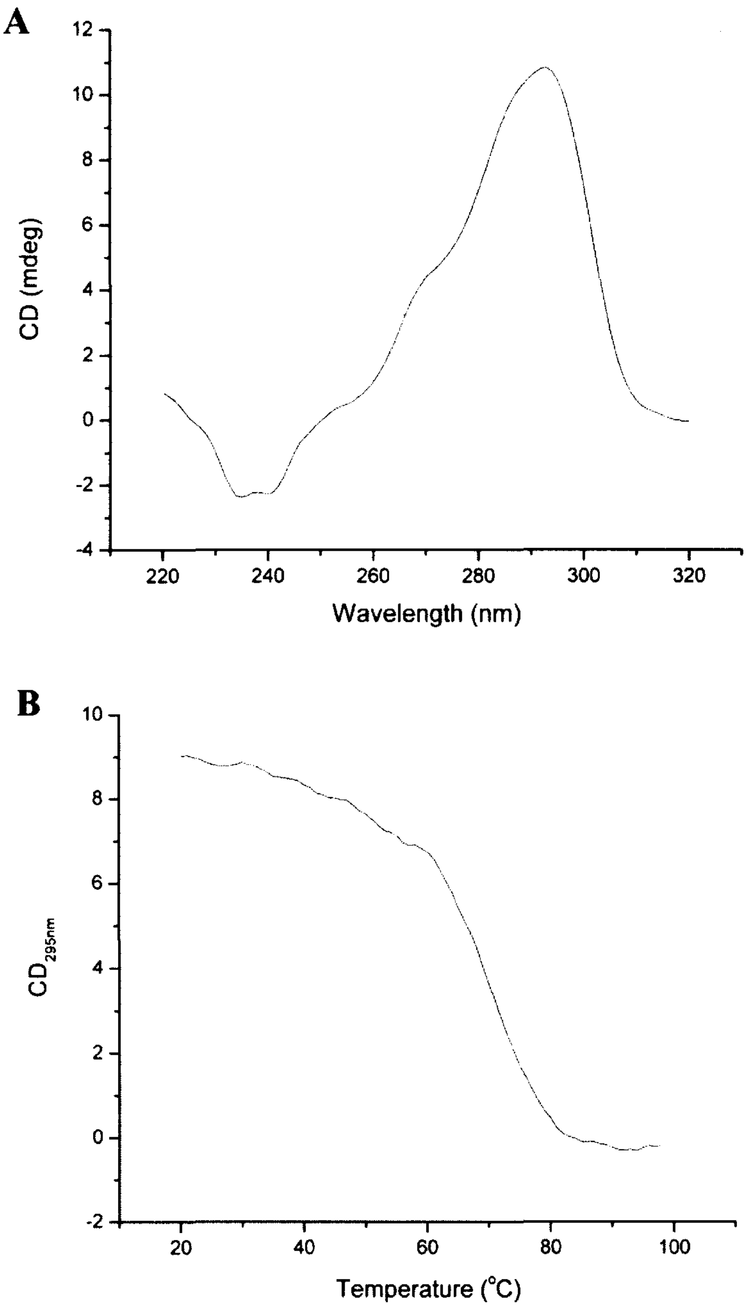
$$RCS = \frac{\sum_{i=1}^{N} Ri}{N} \tag{12}$$

The rank-by-rank strategy (RCS) is used to assign an average rank for each of the top 66,000 compounds from the two available scoring functions in Autodock and Surflex. The following example illustrates how this scoring function works. If the small molecule ranks 1 by Autodock and 3 by Surflex, than the consensus ranking score is 2, using equation 12 above [207]. The RCS is performed for all 66,000 compounds to develop a consensus ranked list of compounds.

Oligonucleotide and Small Molecule Preparation. The G-quadruplex oligonucleotide AGGG(AGGGTT)$_3$, also referred to here as the "K$^+$ 22mer," was obtained from Integrated DNA technologies (IDT, Coralsville, IA) and prepared for experiments by dialysis and annealing. Dialysis was performed against KPEK buffer, which is composed of K_2HPO_4 (6mM), KH_2PO_4 (2mM), KCl (185mM), EDTA (0.1mM), pH 7, using Pierce (Rockford, IL, USA) 3500 Da molecular weight cutoff dialysis cassettes. The oligonucleotides were annealed by heating at 90 ˚C for 2 minutes, cooling to room temperature overnight and left at 4 ˚C for 48 hours prior to use, as previously described [10, 64]. The oligonucleotide (with $\varepsilon = 228{,}500$ L/(mole cm) for the single-strand form) was characterized structurally by Circular Dichroism (CD) spectroscopy and the resulting spectrum was consistent with previously reported results for this structure (Figure 65) [132, 208]. All oligonucleotides used for the quadruplex melting studies were also obtained, annealed and characterized using the methods described and

Figure 65. CD Spectrum (A) and CD Melt (B) of the K^+ 22mer G-Quadruplex. Both spectra are consistent with published structural data for this nucleic acid.

Figure 65. CD Spectrum (A) and CD Melt (B) of the K^+ 22mer G-Quadruplex.



previously [10, 64]. The top nine small molecules with the best ranked consensus scores were purchased for testing. The molecule described in detail here has ZINC identification number 8927810, was purchased from InterBioScreen (Moscow, Russia) under the catalog number STOCK1S-61623 and is described chemically as 1-methyl-4-[5-(1-methyl-4-quinolylidene)-3-phenyl-penta-1,3-dienyl]-quinoline. This small molecule will be hereafter described as "Compound 1." Compound 1 was weighed and a stock solution was created by dissolved the weighed compound in DMSO (Sigma, St. Louis, MO, USA) prior to testing.

Biophysical Testing Methods. UV/Vis Absorption titration experiments were performed on a Tecan Safire 96 well microplate reader (Durham, NC, USA) in duplicate and measured five times at 1nm step intervals between 550nm and 950nm, consistent with previously described procedures [203]. The percent hypochromicity for the UV/Vis absorption titrations was determined from the shift in the absorbance at no added DNA (650nm) and maximal added DNA (659nm) by the formula:

$$Hypochromicity\ (\%) = \frac{Abs650nm - Abs659nm}{Abs650nm} X\ 100 \tag{13}$$

Both the UV/Vis Absorbance and CD experiments were performed using procedures previously described [184, 203, 209]. All CD experiments were performed on a Jasco J-810 spectropolarimeter (Easton, MD, USA). CD scanning experiments for the purposes of K^+ 22 mer quadruplex DNA characterization were performed at a concentration of 3.5μM (strand) from 320nm to 220nm with a data interval of 1nm, band width of 1nm, response of 1 second, scanning speed of 200 nm/minute and a total of four accumulated scans. Induced CD experiments were performed from 900nm to 550nm with a data interval of 1nm, band width of 1nm, response of 2 seconds, scanning speed of 200 nm/minute and a total of three accumulated scans. For induced CD experiments, the Compound 1 concentration was fixed at 11.6μM and separate solutions were prepared with increasing ratios of K^+ 22 mer quadruplex DNA. CD melting experiments with the K^+ 22mer were performed at a DNA concentration of 3.5μM from 20 °C – 98 °C with a scanning speed of 1 °C/minute.

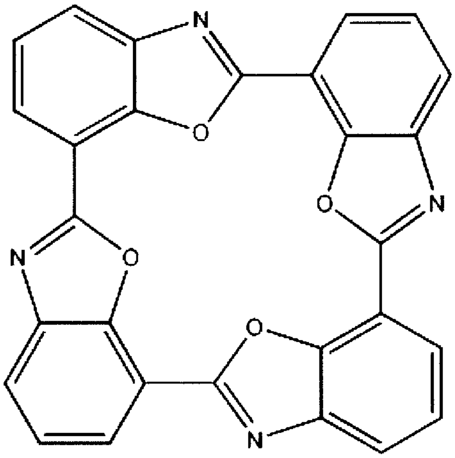
The Thiazole Orange Fluorescent Intercalation Displacement Assay (TO-FID) was performed using procedures previously described [40]. Thiazole Orange was obtained from Anaspec, Inc (San Jose, CA, USA) and dissolved in DMSO. Using a K^+ 22mer quadruplex concentration of 0.25μM (strand) and a Thiazole Orange concentration of 0.50μM, increasing amounts of Compound 1 test ligand are added to the solution and the fluorescence of Thiazole Orange is monitored. All TO-FID fluorescence readings were performed in duplicate and measure 5 times on a Tecan Safire 96 well microplate reader with Excitation at 501nm, Emission from 521nm to 750nm, Emission maximum at 535nm, 1nm step size, Excitation and Emission Band Widths of 9nm and a Gain of 130. All spectroscopic and TO-FID testing was performed in a buffer solution consisting of KPEK and 5% DMSO. The absorbance of Compound 1 is in the region of 550nm – 900nm and does not interfere with reading TO fluorescence at 535nm.

The quadruplex melting studies were performed in a 96-well plate format on an Applied Biosystems StepOnePlus Real-Time PCR System (Carlsbad, CA, USA) adapted for use in thermal melting experiments. The K^+ 22mer quadruplex was labeled with a FAM-TAMRA Fluorescence Resonance Energy Transfer (FRET) pair to selectively monitor the K^+ 22 mer quadruplex melting in the presence of competing DNA solutions. For the FAM-TAMRA labeled K^+ 22mer quadruplex compound melting saturation experiments, increasing concentrations of Compound 1 were added to a fixed solution containing 250 nM FAM-TAMRA labeled K^+ 22mer quadruplex (Sigma, St. Louis, MO, USA) and melting experiments were performed. The temperature range for the melting range was 20°C – 98°C with data measurements taken every 0.2°C. The FAM-TAMRA labeled quadruplex was monitored using a fluorescence filter that quantifies emission at 520 nm. Melting curves were fit in Mathematica v6.0.2.1 (Wolfram Research) and melting temperatures calculated as previously described [210-211]. For the DNA competition experiments, a stock solution of 250 μM of FAM-TAMRA labeled K^+ 22mer quadruplex was made by weighing the quadruplex, dissolving in a solution of tetrabutyl ammonium phosphate and adjusting the pH to 7.0 with tetrabutyl ammonium hydroxide. The stock solution of FAM-TAMRA labeled K^+ 22mer quadruplex was diluted to a final concentration of 150 nM using KPEK buffer for all melting experiments. Competing DNA solutions were added to the wells containing the FAM-TAMRA labeled K^+ 22mer quadruplex such that the final concentration of competing DNA was 20 fold higher (3 μM) than the concentration of the FAM-TAMRA labeled K^+ 22mer quadruplex (150 nM). Finally, Compound 1 was added such that the concentration ratio of Compound 1 to FAM-TAMRA labeled K^+ 22mer quadruplex was 40/1 (final Compound 1 concentration of 6 μM). All melting studies were performed in a buffer consisting of KPEK + 5% DMSO. Oligonucleotides concentrations were based on a monomeric unit (nucleotide for duplex DNA, triplet for triplex DNA and quartet for quadruplex DNA) as previously described [172-173].

Hybrid-1 and Hybrid-2 Quadruplex Modeling and Docking Methods. The nucleic acids that are representative of the Hybrid-1 (PDB ID: 2HY9) and Hybrid-2 (PDB ID: 2JPZ) structures that are present in the K^+ 22mer quadruplex sequence were downloaded from the Protein Data Bank in PDB file format and prepared for molecular docking experiments as described above. Both of these structures contain the human telomeric repeat sequence $(TTAGGG)_n$ with Hybrid-1 consisting of the sequence $AAAGGG(TTAGGG)_3AA$ and Hybrid-2 consisting of the sequence $(TTAGGG)_4TT$. There are four possible end-pasting sites in these two structures to which Compound 1 was docked using both Surflex and Autodock. In each of these two structures, end-stacking sites are present at both the 3' and 5' ends and occupy the space between the terminal guanine tetrad and accompanying loop structures. Unfortunately, Surflex v2.2 has yet to include and have receptor flexibility validated for use with targeting nucleic acids. Therefore, an alternative strategy was employed to "open up" the external G-quartet and surrounding bases that comprise the end-pasting site to allow for molecular docking to proceed. To expose the end-pasting sites, we built a virtual ligand consisting of a quaterpurine, a largely planar, aromatic, small molecule that would stack well upon the terminal guanine quartet. This virtual ligand was appropriately named because it consists of four purines that are connected together in a cyclic arrangement (Figure 66). Using Macromodel, the quaterpurine was initially positioned between the terminal G-tetrad and loop region for each of the possible four endpasting sites. The nucleic acid was initially energetically minimized holding the ligand fixed using a Steepest Descent algorithm for 1000 iterations. The nucleotides comprising the end-pasting site including the terminal G-quartet and loop nucleotides were designated as flexible while the remaining nucleic acid bases were held rigid and fixed. Further structural minimizations were performed by 500 iterations of the Polak Ribier Conjugate Gradient Method. This

Figure 66. The structure of the Quaterpurine small molecule used to “open-up” the Hybrid-1 and Hybrid-2 End Pasting sites.

Figure 66. The structure of the Quaterpurine small molecule used to "open-up" the Hybrid-1 and Hybrid-2 End Pasting sites.



approach successfully opened the end-pasting sites for Surflex and Autodock molecular docking experiments. The docking of Compound 1 to the two end pasting sites in Hybrid-1 and the two end pasting sites in Hybrid-2 using Surflex and Autodock was performed based on our previously published results validating the use of these software to target small molecules that bind to nucleic acids [63]. For Surflex, the "Multistart 5" option was enabled and with Autodock, the number of dockings was set to 5 and the number of energy evaluations was set to 20 million (2E7) [63]. The "protomol" used by Surflex was generated using the position of the quaterpurine ligand occupied in the end pasting sites. The Autodock procedures for grid map preparation and grid parameters used for docking have been previously described (Table 11) [63]. These molecular modeling studies of Compound 1 with the human telomeric structures were performed on a single computer consisting of 2.66GHz Intel(R) Xeon(R) E5430 processor and required an average of approximately 2.2 minutes and 31.8 minutes to complete each of the four Surflex and Autodock experiments, respectively.

Results and Discussion

The (TTAGGT)$_4$ Quadruplex has a Representative End-Pasting Site for Virtual Screening. The discovery of small molecules that are able to bind to the end-pasting region of the human telomeric sequence AGGG(TTAGGG)$_3$ (K$^+$ 22mer) are of great interest as they can potentially stabilize the quadruplex *in vivo* and may possess anti-cancer activity [55]. A search of the online RSCB PDB database showed over 120 quadruplex structures that had been deposited as of December 2009. Surprisingly, however, there are conflicting published reports about the "correct" structure of the K$^+$

Table 11. Autodock Docking Parameters used for *In silico* Targeting of the End-Pasting Sites in Hybrid-1 and Hybrid-2.

Table 11. Autodock Docking Parameters used for *In silico* Targeting of the End-Pasting Sites in Hybrid-1 and Hybrid-2.

Autodock Target	**Dimensions of Grid (X, Y, Z)**	**Grid Center**
Hybrid 1 End Paste 1	66 X 46 X 70	-0.034 X 7.176 X 0.032
Hybrid 1 End Paste 2	70 X 64 X 66	-0.01 X 4.819 X 0.036
Hybrid 2 End Paste 1	50 X 40 X 50	-0.099 X 7.303 X 0.059
Hybrid 2 End Paste 2	68 X 38 X 56	0.089 X 3.91 X -0.018

22mer under physiologically relevant conditions, even though there are multiple small molecules that are thought to interact with the human telomeric quadruplex by end-pasting binding [191, 202]. Increasing evidence is suggesting that the K^+ 22mer exists as a "hybrid" structure, but there have yet to be any NMR structures published with small molecules bound to this structure [191, 193-194]. This is unfortunate, as the ligand could be easily removed computationally and the nucleic acid used for structure-based virtual screening. Instead, however, structures that possess the human telomeric repeat must be computationally altered prior to molecular docking to allow docking to the end-pasting region of the quadruplex.

An alternative approach that we propose is selection of a G-quadruplex structure (with a small molecule complexed) which we believe possesses similar properties of the K^+ 22mer end-pasting including a terminal guanine quartet with flanking Adenine containing residues. Interestingly, the quadruplex $(TTAGGGT)_4$ with PDB ID 1NMZ appears to possesses many properties that makes it a representative end-pasting site of the K^+ 22mer G-quadruplex structure. While the small molecule that is found in the $(TTAGGGT)_4$ quadruplex site lies in an end-pasting site with adenine and guanine quartets (Figure 64B), we believe that this arrangement of nucleotides is remarkably similar to an end-pasting site that is present in the K^+ 22mer (Figure 64A) and thus makes $(TTAGGGT)_4$ a suitable choice for virtual screening experiments. The small molecule RHPS4 that is found in the end-pasting site of $(TTAGGGT)_4$ is easily removed computationally and allows the nucleic acid structure to be used with minimal alterations for virtual screening experiments. In this sense, the $(TTAGGGT)_4$ structure represents the *in silico* structure that is perhaps singularly most representative of an end-pasting site found in the K^+ 22mer. We prefer this strategy over previously reported approaches which typically involve breaking bonds of the target nucleic acid structure to generate an artificial site for docking as our approach preserves the fidelity of the *in silico* structure without manipulation or perturbation [212]. Importantly, this also demonstrates an approach for circumventing structure-based *in silico* screening problems when the target site is not entirely available from the *in silico* structure.

Virtual Screening Discovery of a Novel Quadruplex Binding Ligand. The $(TTAGGGT)_4$ structure with the representative pseudo end-pasting site was the basis for the virtual screening efforts to discover new quadruplex binding small molecules. The next challenge is the selection of molecular docking programs that are suitable for screening millions of compounds against a nucleic acid target. Our recent validation of two popular protein molecular docking programs, Autodock and Surflex, for use with nucleic acids demonstrated that these software can accurately reproduce multiple small molecule crystal structures of small molecules that interact with DNA by several mechanisms [34, 63]. The docking results showed that while both docking programs were accurate at reproducing the crystal structures, Autodock required approximately 10 fold greater time for docking experiments compared to Surflex, under conditions reported previously [63]. This difference in docking speed is particularly important when the *in silico* library screened here is greater than 6.6 million small molecules. The question remains whether these two structure-based docking software can be combined in an *in silico* platform to target the $(TTAGGGT)_4$ pseudo end-pasting site.

While one possibility for virtual screening was to perform experiments using only Surflex, this would neglect the validated use of Autodock, as well as the fact that

Autodock implements different scoring and docking algorithms than Surflex, which may complement the Surflex approach [60, 205]. An alternative strategy that we developed is to use Surflex to pre-screen the entire *in silico* library and subsequently perform Autodock docking on only the top 1% of ranked Surflex hits (Figure 67). This approach balances both the computational efficiency and accuracy of Surflex with the power of Autodock in a single, integrated, virtual screening platform. The top ranked hits (66,000 compounds) for both programs were subsequently re-ranked using a rank-by-rank consensus scoring function that is preferable over a single scoring function and has previously been shown to increase success with virtual screening experiments [206-207]. In this case, a consensus approach is preferred as there is yet to be developed a "universal" scoring function that is suitable for either nucleic acids or proteins. Because each scoring function has distinct advantages and disadvantages, the adoption of a consensus-based approach increases the probability of discovering novel small molecules, while minimizing false positives that commonly occur in virtual screening studies of large numbers of *in silico* compounds [207, 213]. The virtual screening approach outlined in Scheme 1 was used for the selection of nine small molecules with hypothesized K^+ 22mer end-stacking binding. Unfortunately, eight of the nine compounds were excluded from biophysical testing in the assays described herein due to such problems as solubility limitations or lack of a suitable chromophore for testing. However, one small molecule in particular, Compound 1 (Figure 68), possessed suitable properties for biophysical experiments and was tested to determine if the hypothesized activity that was identified through the *in silico* screen could be demonstrated *in vitro*.

Figure 67. The virtual screening strategy used to dock the *in silico* library of compounds to the $(TTAGGT)_4$ quadruplex and determine the best hits for biophysical testing.

Figure 67. The virtual screening strategy used to dock the *in silico* library of compounds to the $(TTAGGT)_4$ quadruplex and determine the best hits for biophysical testing.

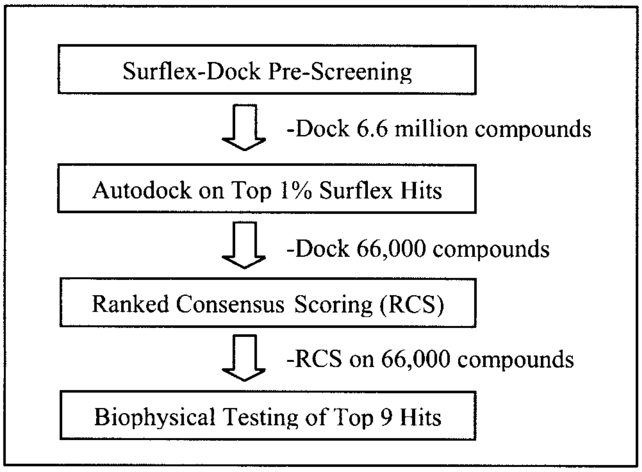
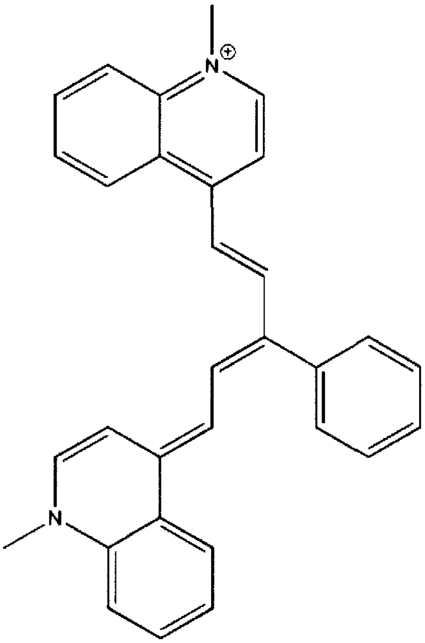

Surflex-Dock Pre-Screening
-Dock 6.6 million compounds
Autodock on Top 1% Surflex Hits
-Dock 66,000 compounds
Ranked Consensus Scoring (RCS)
-RCS on 66,000 compounds
Biophysical Testing of Top 9 Hits

Figure 68. The newly discovered small molecule, Compound 1, from the *in silico* virtual screening experiments.

Figure 68. The newly discovered small molecule, Compound 1, from the *in silico* virtual screening experiments.



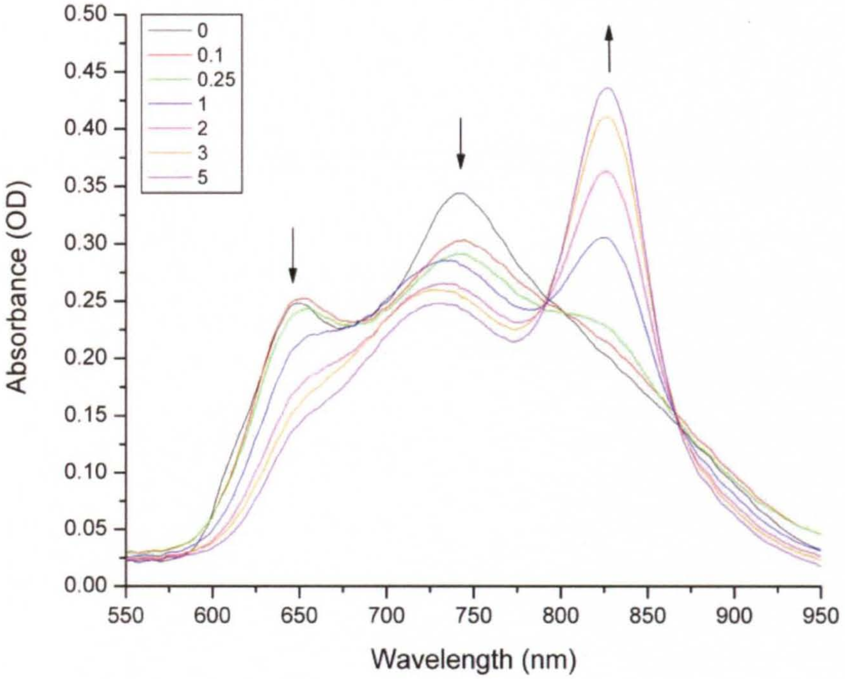
These experiments were crucial to determine if the virtual screening strategy developed in Scheme 1 can not only discover new compounds, but also predict the specific site of binding of the compound to the quadruplex.

Biophysical Validation of the Predicted In silico Activity of Compound 1. While our *in silico* approach successfully discovered multiple ligands with hypothesized K^+ 22 mer G-quadruplex binding activity, the question remains whether the computational methods are truly predictive of small molecule nucleic acid interactions *in vitro*. To address this concern, we present biophysical testing that suggests that our molecular docking approach as outlined in Scheme 1 successfully identifies a small molecule that not only interacts with the K^+ 22mer quadruplex as predicted *in silico*, but also binds by the hypothesized end-stacking mechanism to the external guanine quartet.

UV/Vis Absorbance Titrations and CD Spectroscopy. UV/Vis absorption titrations and Circular Dichroism (CD) are powerful techniques that for our purposes can be used to confirm our *in silico* predictions of the interaction of Compound 1 with the K^+ 22mer. The UV/Vis absorption experiments involve adding an increasing amount of K^+ 22mer DNA to a solution of fixed ligand concentration while monitoring the absorbance spectrum of the ligand. Because the absorption spectrum of the ligand (550 nm – 900 nm) is unique from that of the DNA (<300 nm), changes in the monitored ligand spectra are indicative of specific interactions with the DNA. These spectra data show unambiguously that Compound 1 binds to the K^+ 22mer. In the case of Compound 1, as increasing concentrations of K^+ 22mer DNA are added to the solution, there is marked hypochromicity that occurs at 650nm as well as an appearance of a new peak at 827nm (Figure 69). The amount of hypochromicity and relative wavelength shift can be used as

Figure 69. UV/Vis Absorption Titrations at increasing Quadruplex/Compound 1 ratios demonstrating the G-quadruplex interaction of Compound 1 with the K^+ 22mer quadruplex, AGGG(TTAGGG)$_3$.

Figure 69. UV/Vis Absorption Titrations at increasing Quadruplex/Compound 1 ratios demonstrating the G-quadruplex interaction of Compound 1 with the K^+ 22mer quadruplex, AGGG(TTAGGG)$_3$.

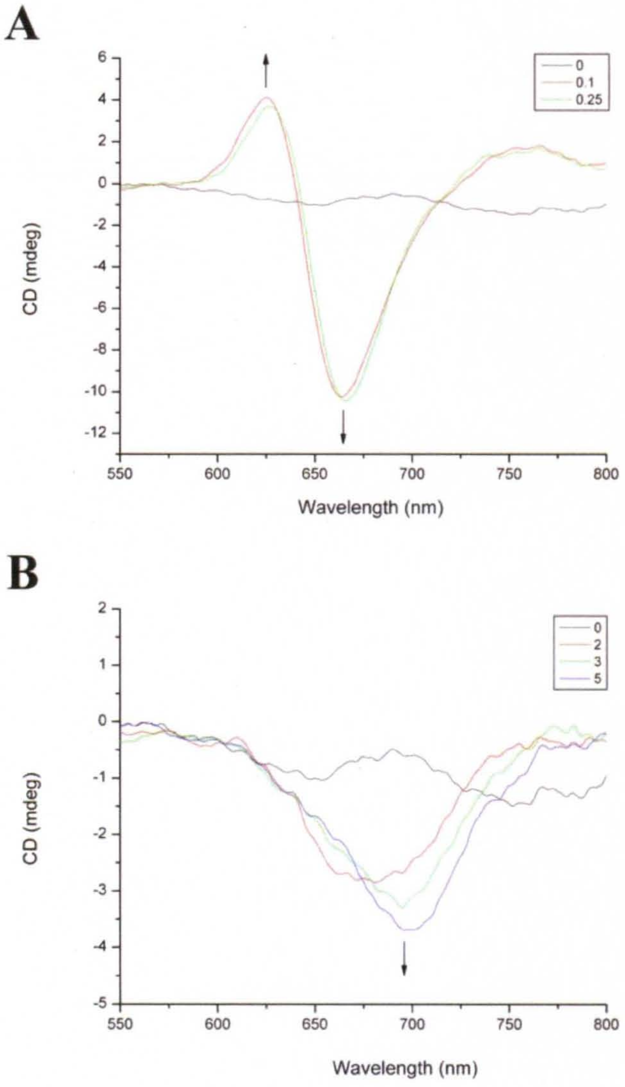


indicators to determine the DNA binding mode [214]. The amount of hypochromicity observed here (>38.2%) is consistent with ligands such as TmPyP4 that can interact with the K^+ 22mer by both end-stacking and intercalation [132]. However, intercalation of ligands such as TmPyP4 is typically characterized by a bathochromic, red shift of > 15 nm, while reports for ligands that interact by end-stacking such as Berberine are in the range of 8 – 12 nm [132, 146, 203, 209, 215]. Additionally, it has been reported that intercalation is energetically less favorable than terminal end-stacking due to the challenge of quadruplexes accommodating ligands stacked between existing guanine quartets [209]. In the case of Compound 1, in addition to the observed hypochromicity at 650 nm, an estimated 9 nm red shift occurs from 650 nm to approximately 659 nm, suggesting that Compound 1 may interact with the K^+ 22mer quadruplex by end-stacking. A consideration here is that the amount of hypochromicity and red-shift appears to be somewhat ligand-and quadruplex dependent, and the amount to which the quadruplex loop interactions are involved can substantially impact these values [209]. Nonetheless, the UV/Vis absorption data support an interaction of Compound 1 with the K^+ 22mer quadruplex and suggest an end-stacking interaction with the external G-tetrads of the K^+ 22mer quadruplex.

Circular Dichroism was also employed to determine if Compound 1 interacts with the K^+ 22mer quadruplex by the *in silico* predicted end-stacking mode. Circular Dichroism is useful for studying small molecule-nucleic acid interactions as ligands typically lack a CD signal, but upon binding to DNA, an "induced" CD (ICD) effect may be observed. Importantly, the magnitude and sign of the ICD signal typically allows for the classification of the ligand binding mechanism as either groove binding (a positive signal) or intercalation (a negative signal) [184]. The CD results are consistent with the other data presented thus far and clearly show an interaction of Compound 1 with the K^+ 22mer quadruplex. At low Quadruplex/Compound 1 ratios (0.1 – 0.25), Figure 70A shows the presence of both a positive peak in the region of 625nm and a negative peak in the region of 665nm. The spectroscopic signature, with the presence of bisignate positive and negative peaks, is likely the so-called "exciton" effect. The exciton effect is significant as it may indicate the presence of multimers or aggregates of Compound 1 that impart this unique spectroscopic signal [184]. At higher Quadruplex/Compound 1 ratios (2 – 5), there is a progressive negative inducible CD that is observed as well as shift from approximately 675 nm - 700 nm. This behavior is generally consistent with Compound 1 interacting with the quadruplex by intercalation or end-stacking (Figure 70B). The shift of the CD spectral minimum from 675 to 700 nm could exist for several reasons. First, there may exist multiple end-pasting sites with variable affinity for the ligand. This site-dependent variable affinity binding behavior has been seen previously with other K^+ 22mer binding small molecules such as TmPyP4 [203]. Additionally, because the K^+ 22mer DNA sequence typically consists of a heterogeneous mixture of species, there may be interactions with multiple, unique G-quadruplex end-pasting sites [194]. It is also possible that the ligand may interact primarily by end-stacking and by intercalation, as this dual type of binding behavior has been previously reported with other G-quadruplex binding ligands [131]. In total, however, the appearance of an induced CD signal as seen here is consistent with other biophysical data and supports the end-stacking of Compound 1 to the K^+ 22mer quadruplex.

Figure 70. Induced CD Spectroscopy at increasing Quadruplex/Compound 1 ratios showing the G-quadruplex interaction of Compound 1 with the K^+ 22mer quadruplex, AGGG(TTAGGG)$_3$. (A) and (B) show the spectroscopic profile at lower and higher Quadruplex/Compound 1 ratios, respectively.

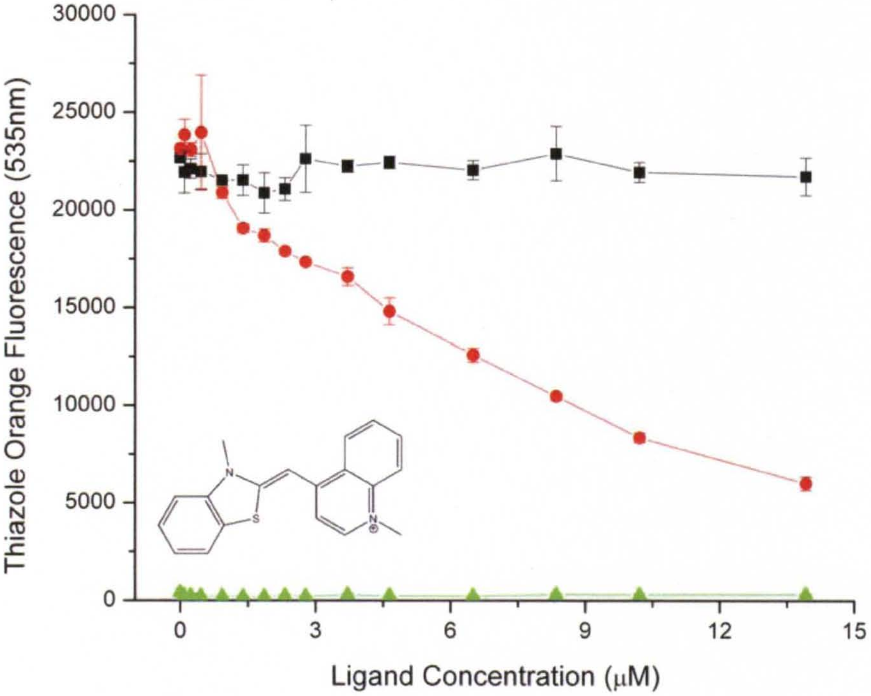
Figure 70. Induced CD Spectroscopy at increasing Quadruplex/Compound 1 ratios showing the G-quadruplex interaction of Compound 1 with the K^+ 22mer quadruplex, $AGGG(TTAGGG)_3$. (A) and (B) show the spectroscopic profile at lower and higher Quadruplex/Compound 1 ratios, respectively.



Thiazole Orange Fluorescence Intercalation Displacement (TO-FID). The TO-FID assay is a complementary assay to the UV/Vis absorbance and CD spectroscopic studies and is another method to determine if Compound 1 interacts with the K^+ 22mer quadruplex by end-stacking. One of the advantages of this assay is that only the fluorescence of thiazole orange (TO) is monitored, so the tested ligand need not have any fluorescent or absorbance properties. The assay requires no specialized equipment and less training than other techniques that have been used to assess DNA-small molecule interactions such as ESI-MS and SPR [40]. Finally, the assay is amenable to a 96 well microplate format; a property that is desirable for high-throughput testing of ligands that may be discovered by virtual screening and other computational methods. Successful operation of the TO-FID assay relies on the known binding of TO to the end-pasting region of the human telomeric quadruplex K^+ 22mer which has been previously well characterized [40, 216-217]. Thiazole Orange is unique as it typically has a reported several hundred to thousand fold increase in fluorescence when it is bound to DNA and an insignificant fluorescence when displaced from DNA and unbound in solution [217-219]. The assay is performed by initially saturating the K^+ 22mer DNA with TO and subsequently adding increasing amounts of the ligand to the solution. If the ligand binds to the quadruplex end-stacking region with sufficiently high affinity, TO will be displaced and the fluorescence will be quenched. This assay can thus confirm both binding of the ligand to the quadruplex and the probable site of binding. Indeed, the quenching of TO fluorescence upon addition of Compound 1 suggests that it dissociates TO from the end-pasting region of the K^+ 22mer (Figure 71). The testing of Compound 1 was successful and avoided the limitations of the assay, which are as follows. First, if the

Figure 71. TO-FID results showing the displacement of Thiazole Orange (TO) from the K^+ 22mer quadruplex and subsequent TO fluorescence quenching at increasing concentrations of the ligand Compound 1. The inset shows the structure of TO. Black squares = TO + DNA + buffer. Red circles = TO + DNA + Compound 1. Green triangles = TO + Compound 1. Data plotted are the average of experiments in duplicate. Error bars represent ± one standard deviation.

Figure 71. TO-FID results showing the displacement of Thiazole Orange (TO) from the K^+ 22mer quadruplex and subsequent TO fluorescence quenching at increasing concentrations of the ligand Compound 1. Data plotted are the average of experiments in duplicate. Error bars represent ± one standard deviation.

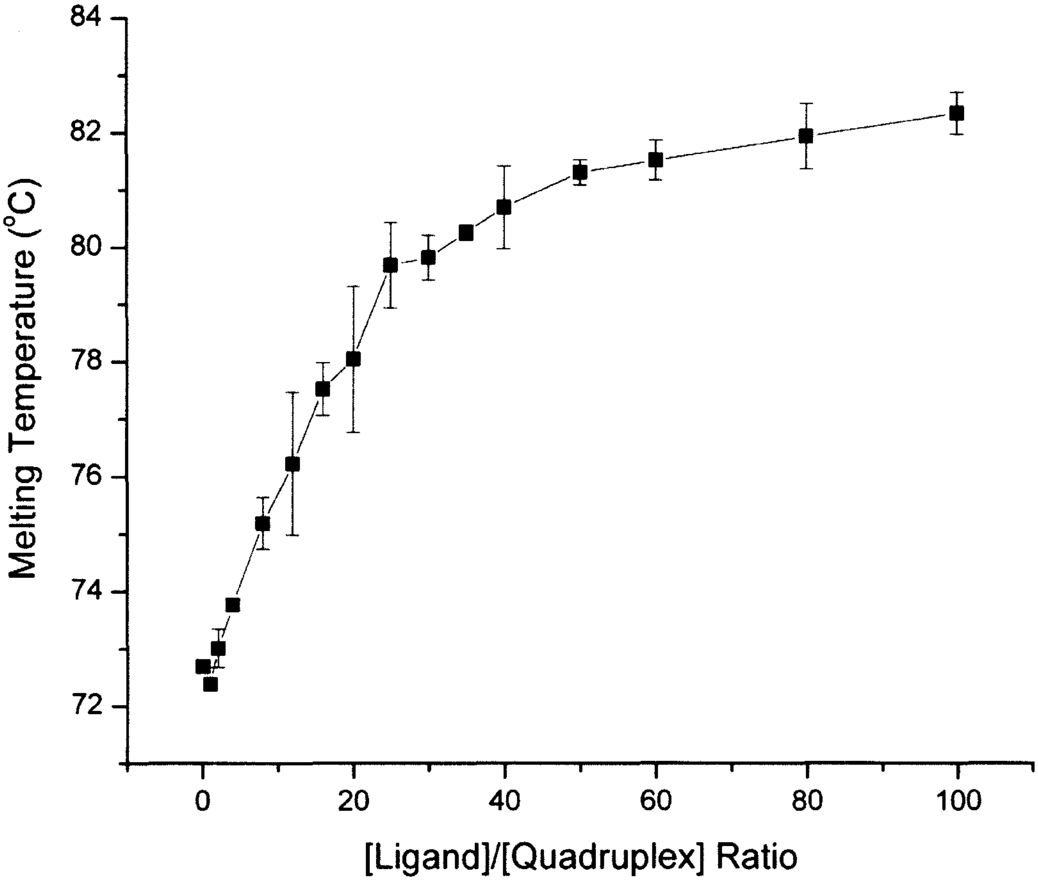


ligand is of weak affinity (or has a binding constant lower than that of TO), it may not be sufficient to displace TO. Second, the ligand may bind the quadruplex at a different site, the grooves for example, instead of the end sites, and this binding would elude detection by the FID method as described here. Third, the ligand of interest may have overlapping spectroscopic behavior with Thiazole Orange which can confound the results of the assay. Compound 1 circumvents these limitations as it clearly displaces and lacks any spectral overlap with the TO probe, allowing an assessment of the binding properties of Compound 1. It is conceivable that Compound 1 may also interact with the K^+ 22mer by additional mechanisms at distinct sites on the K^+ 22mer that may complicate the observed fluorescent quenching curve. However, in total, the TO-FID data are consistent with the UV/Vis and CD data and support the initial *in silico* based hypothesis that Compound 1 likely interacts with the end-pasting sites on the human telomeric K^+ 22mer quadruplex.

Selective Recognition and Thermal Stabilization of the K^+ 22mer Quadruplex. Melting experiments were performed to determine whether Compound 1 can thermally stabilize the K^+ 22mer quadruplex in the absence and presence of competing DNA structures. Many of the most studied quadruplex binding ligands have been shown to increase the quadruplex melting temperature (T_m) [220]. It is worth mentioning that the melting temperature of the K+ 22mer quadruplex (as tested here) occurs at approximately 72°C which is far above physiological temperature (37°C) and may not reflect what occurs under more biologically representative conditions [221]. However, the assay provides further unambiguous proof of binding of a compound to the quadruplex target. As Figure 72 shows, the addition of increasing amounts of ligand to the K^+ 22mer results

Figure 72. Effect of Compound 1 on the Melting Temperature of the FAM-TAMRA labeled K^+ 22mer G-Quadruplex.

Figure 72. Effect of Compound 1 on the Melting Temperature of the FAM-TAMRA labeled K$^+$ 22mer G-Quadruplex. Data plotted are the average of experiments performed in triplicate. Error bars are ± one standard deviation.



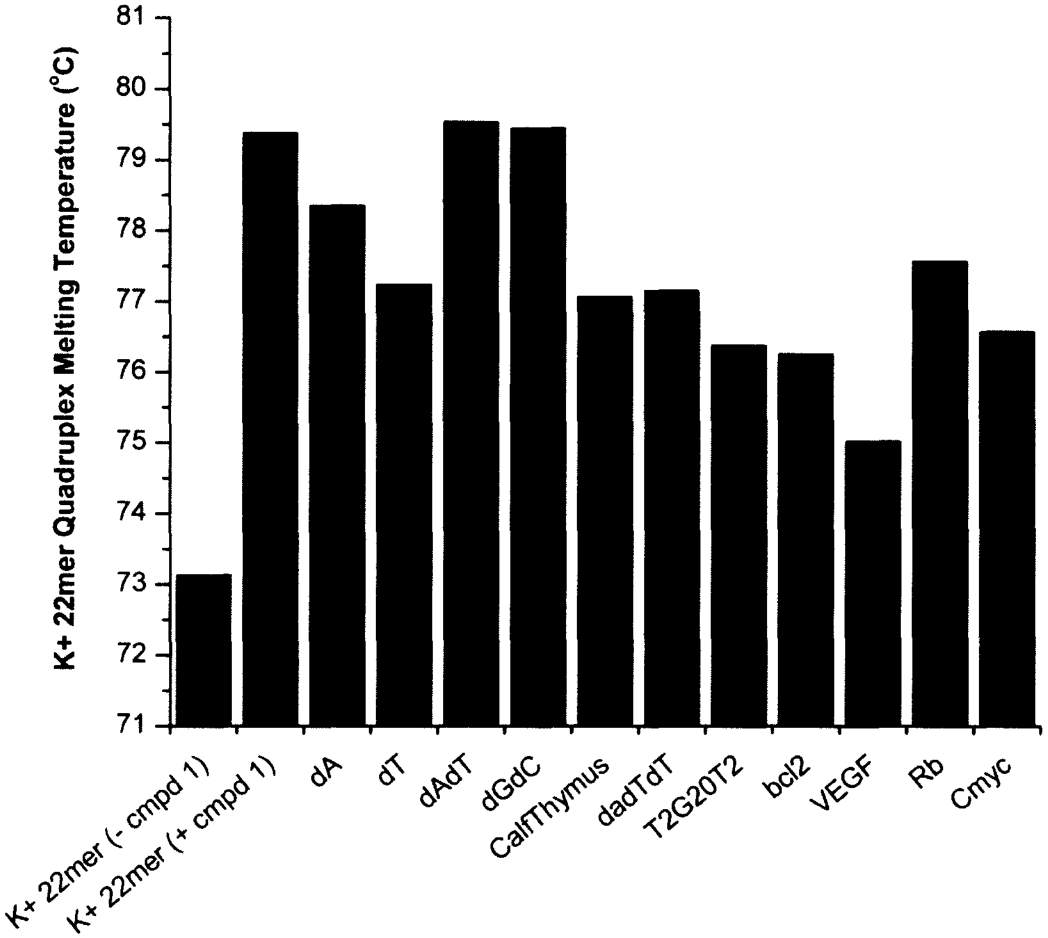
in an increase in the T_m of the quadruplex of approximately 10°C at the highest levels of Compound 1 tested.

Competition experiments were performed to determine if Compound 1 prefers the K^+ 22mer over other DNA structures. Under the conditions tested, Compound 1 stabilizes the K^+ 22mer from approximately 73°C (– cmpd 1) to 79.5°C (+ cmpd 1). The addition of 20 fold excess competitor quadruplex ($T_2G_{20}T_2$, bcl-2, VEGF, Rb, C-myc) relative to the K^+ 22mer can generally decrease the T_m of the K^+ 22mer which suggests that these structures are successfully competing for binding to Compound 1 (Figure 73). The VEGF quadruplex appears to be most successful at competing for Compound 1 as is evidenced by the greatest decrease in K^+ 22mer T_m. However, the addition of dAdT and dGdC does not effect the T_m, which suggests that Compound 1 selectively prefers the K^+ 22mer quadruplex compared to these DNA structures. Taken in total, these results further validate the binding of Compound 1 to the K^+ 22mer quadruplex and also suggest that Compound 1 offers some selectivity for quadruplex compared to duplex DNA structures.

Molecular Docking of Compound 1 to Hybrid-1 and Hybrid-2. With the initial discovery of Compound 1 and subsequent biophysical testing to demonstrate binding activity to the K^+ 22mer G-quadruplex, the next goal was to develop possible models for how Compound 1 can interact with the external G-quartet that is present in the K^+ 22mer quadruplex. These models could provide valuable insights into the possible interaction of Compound 1 with the terminal G-quartets and loop regions that are present in the human telomeric quadruplex structure. The K^+ 22mer is polymorphic and is known to adopt several so-called Hybrid structures (Hybrid-1 and Hybrid-2) *in vitro* [146].

Figure 73: Effect of a 20 fold Excess of Competitor DNA on the Ability of Compound 1 to Thermally Stabilize the FAM-TAMRA labeled K^+ 22mer G-Quadruplex. Duplex, triplex and quadruplex nucleic acids were used as competitor DNA structures.

Figure 73: Effect of a 20 fold Excess of Competitor DNA on the Ability of Compound 1 to Thermally Stabalize the FAM-TAMRA labeled K$^+$ 22mer G-Quadruplex. Duplex, triplex and quadruplex nucleic acids were used as competitor DNA structures.



Structures for the Hybrid-1 (PDB ID 2HY9) and Hybrid-2 (PDB ID 2JPZ) quadruplexes are available *in silico* and were selected for modeling studies as they contain the human telomeric K^+ 22mer repeat [196-197, 222], but have unique flanking sequences. These two structures each have two external G-quartet end-stacking sites for a total of four possible end-pasting sites. Interestingly, while the common conserved feature among these sites is the presence of the G-quartet, each of these sites has unique loop topologies that may confer ligand-binding specificity. Unfortunately, a limitation of these structures is that neither of these structures has a complexed ligand present in the end-pasting zone. An alternative approach is to computationally "open-up" the end-pasting site prior to docking of Compound 1. This was accomplished by constructing a planar, aromatic small molecule ligand, manually placing the ligand in the four possible end-pasting sites and energetically minimizing the nucleic acid structure around the ligand (Figure 66). This procedure was sufficient to open up the end-pasting site for molecular docking for the purposes of modeling the fit of Compound 1 in the putative end-pasting sites.

Molecular docking of Compound 1 to all four end-pasting sites was successfully performed with Autodock and Surflex using a validated docking procedure previously described for nucleic acids [63]. The top poses of Compound 1 for these docking software are shown in Figure 74 with the resulting scores shown in Table 12. Based on our previous experience with these programs, the Autodock and Surflex scores for all docked complexes are comparable. Generally, Compound 1 appears to dock and fit well in all four of the end-pasting sites. A closer view of the top poses without the nucleic acid present (Figure 75) reveals that the top ranked poses for Surflex and Autodock appear to have a higher amount of overlap for the two Hybrid 1 end-pasting sites

Figure 74. The top ranked poses for Compound 1 docked using Surflex (Gold) and Autodock (Cyan) to the following G-quadruplex nucleic acid structures: (A) Hybrid-1 End Paste Site 1 (B) Hybrid- 1 End Paste Site 2 (C) Hybrid-2 End Paste Site 1 and (D) Hybrid-2 End Paste Site 2. The nucleic acid structures are shown in white except for the small molecule interacting terminal guanine tetrads which are shown in green.

Figure 74. The top ranked poses for Compound 1 docked using Surflex (Gold) and Autodock (Cyan) to the following G-quadruplex nucleic acid structures: (A) Hybrid-1 End Paste Site 1 (B) Hybrid- 1 End Paste Site 2 (C) Hybrid-2 End Paste Site 1 and (D) Hybrid-2 End Paste Site 2.

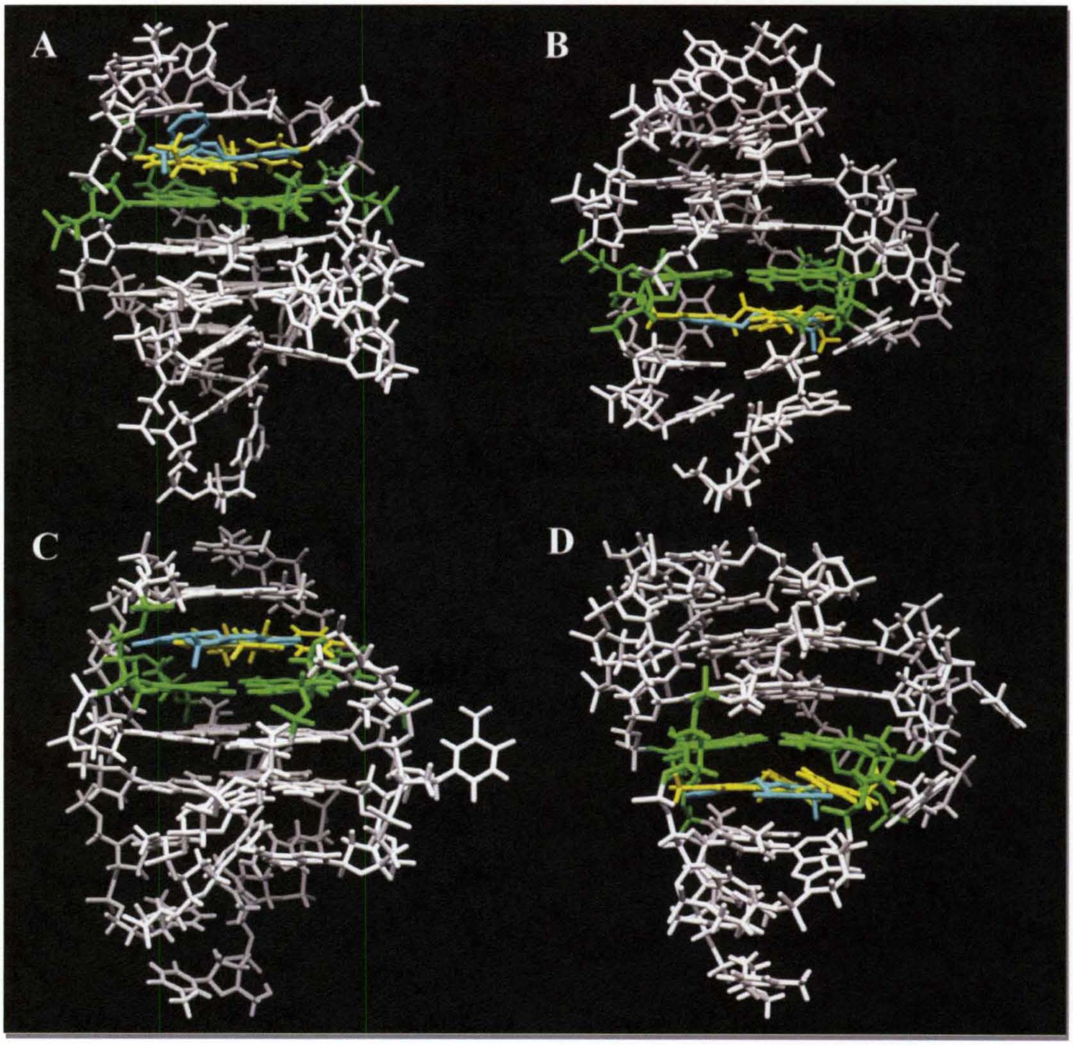


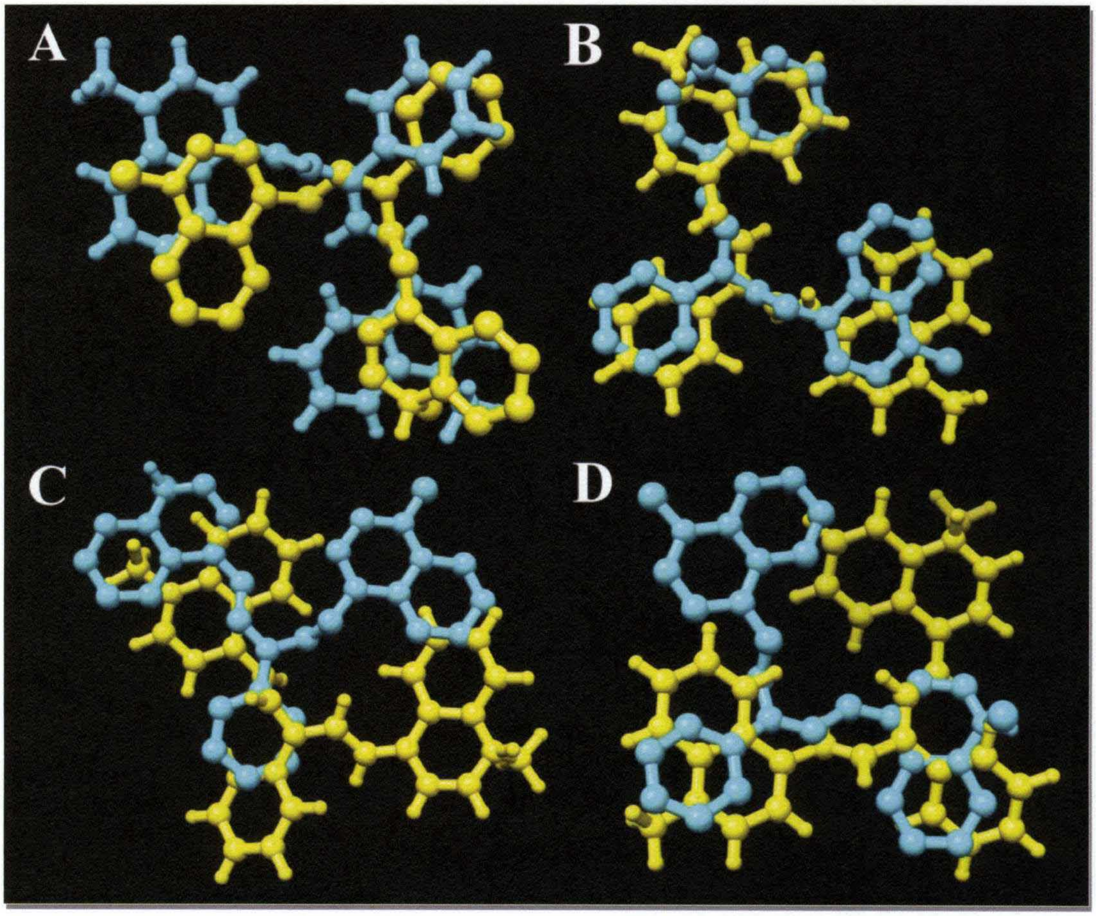
Table 12. The scores for the top-ranked pose from Autodock or Surflex for the four possible end-pasting sites in Hybrid-1 and Hybrid-2.

Table 12. The scores for the top-ranked pose from Autodock or Surflex for the four possible end-pasting sites in Hybrid-1 and Hybrid-2.

Hybrid Structure / PDB ID	**Surflex-Dock Score (-log(K_d))**	**Autodock Score (kcal/mol)**
Hybrid-1 / 2HY9 End Paste site 1	14.29	-11.45
Hybrid-1 / 2HY9 End Paste site 2	15.19	-11.77
Hybrid-2 / 2JPZ End Paste site 1	14.98	-12.71
Hybrid-2 / 2JPZ End Paste site 2	13.56	-12.57

Figure 75. The top ranked poses for Compound 1 docked using Surflex (Gold) and Autodock (Cyan) to the following G-quadruplex nucleic acid structures: (A) Hybrid-1 End Paste Site 1 (B) Hybrid-1 End Paste Site 2 (C) Hybrid-2 End Paste Site 1 and (D) Hybrid-2 End Paste Site 2. The file format for the Autodock poses is PDBQT with merging of non-polar hydrogens.

Figure 75. The top ranked poses for Compound 1 docked using Surflex (Gold) and Autodock (Cyan) to the following G-quadruplex nucleic acid structures: (A) Hybrid-1 End Paste Site 1 (B) Hybrid-1 End Paste Site 2 (C) Hybrid-2 End Paste Site 1 and (D) Hybrid-2 End Paste Site 2.



compared to Surflex and Autodock top ranked poses for the end-pasting sites of Hybrid-2. Visual inspection of the top ranked pose is necessary to evaluate ligand interactions with the guanine quartet and the loop regions. Indeed, in all four of the end-pasting models, Compound 1 appears to stack well on the guanine quartet. This type of binding behavior is in agreement and consistent with that of many largely aromatic compounds such as TMPyP4 that bind G-quadruplex structures by either end-pasting or intercalation [202]. What is typically more difficult to interpret, but albeit equally important, is the interaction of Compound 1 with the various terminal loop structures. In contrast to the Guanine tetrad, the loop arrangements appear much more flexible and conformationally heterogeneous. The models certainly suggest that Compound 1 can interact with the various loops to different extents, which could potentially modulate selective binding and stabilization of quadruplex structures. In fact, compounds that bind by end-stacking as well as preferentially to specific quadruplex loops and accompanying grooves are currently a source of great interest as this is thought to confer quadruplex discriminatory capability to small molecules. Of interest here is the modeling data suggesting both guanine quartet interactions as well as potential loop interactions with the four possible end-pasting sites of the Hybrid-1 and Hybrid-2 quadruplex structures. These models also suggest that Compound 1 may serve as a lead compound for derivatization experiments to increase preferential loop and or groove binding to impart specific quadruplex discrimination.

Conclusions

The discovery of novel G-quadruplex nucleic acid small molecules using *in silico* virtual screening was investigated. This work details the development of an *in silico* virtual screening approach using the end-pasting site from the quadruplex $(TTAGGGT)_4$ as a representative pseudo end-pasting site for the K^+ 22mer quadruplex. The software Surflex and Autodock were used to dock over 6.6 million small molecules to the pseudo end-pasting site and resulted in the discovery of a novel G-quadruplex binding small molecule with a completely novel scaffold. Biophysical testing by spectroscopic and fluorescence based assays validated our virtual screening approach and demonstrated that the strategy was predictive and capable of discovering small molecules that bind quadruplexes specifically by an end-pasting mechanism. Furthermore, four molecular models were developed demonstrating the interaction of the newly discovered compound with the guanine tetrad and loop regions of the end-pasting sites present in structures containing the K^+ 22mer. These results suggest that the *in silico* platform presented here can be utilized to discover new small molecules that have G-quadruplex binding activity and may serve as lead compounds for further modification to optimize quadruplex binding and discriminatory properties.

CHAPTER VI

CONCLUSIONS

The use of molecular docking and virtual screening has gained widespread use for the discovery of novel small molecules for targeting proteins. However, these computational approaches have historically neglected the targeting of nucleic acid structures. With the increase in the understanding of nucleic acid structure and *in vivo* function, nucleic acids have garnered increased attention as relevant and good targets. This is particularly true of the G-quadruplex nucleic acids which can inhibit telomerase activity *in vivo* and decrease cancer cell proliferation. While small molecules have been discovered that bind to G-quadruplexes, many such as TmPyP4 suffer from poor selectivity and bind to many other nucleic acid structures. With the general landscape of "druggable" targets expanding to include nucleic acids such as the G-quadruplexes, there is a critical need to determine if computational resources can be utilized and customized for the discovery of new small molecules that interact with nucleic acid targets selectively and by a known binding mechanism.

The initial goal of this research was determining if the docking software Autodock and Surflex can be used for targeting nucleic acids. We demonstrated conclusively in Chapter II that both software packages can accomplish this goal. The next focus was to determine if *in silico* rules can be developed to predict the mechanism of action and structural selectivity of small molecules that are known to interact with nucleic acids. The answer was affirmative, and *in silico* rules for distinguishing intercalator from groove binders were obtained. This was largely based on the rationalization of known competition dialysis binding data of several sets of small molecules to an array of heterogeneous nucleic acid structures. Finally, the knowledge from this early work was utilized for the discovery of novel triplex and quadruplex binding compounds that may have *in vivo* significance. Importantly, the predicted *in silico* binding behavior of the newly discovered triplex and quadruplex binding compounds was rigorously validated using a combination of spectroscopic, calorimetric and thermodynamic assays. This demonstrates the practical application of the research that is described in this work. This work also provides, for perhaps the first time, an integrated, computational and experimental platform for drug discovery for nucleic acids.

The first set of experiments involved the validation of the molecular docking software Autodock and Surflex for targeting nucleic acid structures. This was accomplished by selecting several minor groove binders (distamycin and pentamidine) and intercalators (daunorubicin and ellipticine) as these ligands have solved *in silico* nucleic acid – ligand structures. Using molecular docking techniques, the software were found to be able to successfully reproduce the *in silico* complexes to a high degree of accuracy. Surflex was discovered to be of comparable accuracy to Autodock but approximately an order of magnitude faster for the molecular docking experiments, which made this software particularly relevant for virtual screening applications. The "optimal" software parameters for virtual screening were determined which served as the basis for the use of these software for the remaining work in this dissertation.

After the establishment of the feasibility of using Autodock and Surflex to target nucleic acids, the next focus was on whether *in silico* rules could be developed to specifically predict the mechanism of action and nucleic acid sequence and structural selectivity of small molecules. Using the four small molecules from the initial validation study (daunorubicin, distamycin, ellipticine and pentamidine), molecular docking experiments were performed for each of the small molecules against an *in silico* array of nucleic acids. Based on these results, *in silico* metrics were developed for Autodock and Surflex to classify each of the small molecules on the basis of their binding mechanism (groove binder or intercalator) and nucleic acid structural and sequence preference. The *in silico* rules were further tested and validated on multiple triplex and quadruplex binding compounds that our lab has discovered as well as an extensive 67 compound library set of compounds for which detailed competition dialysis data exists using the identical array of nucleic acids used for the *in silico* molecular docking experiments. The results supported the use of the metrics for generally successfully predicting the mechanism of action of ligand. Prediction of sequence and structural selectivity of the small molecules generally appeared to be more challenging, especially for some of the larger molecules that were tested.

The development of the *in silico* metrics set the stage for application of the metrics in large scale virtual screening experiments for the discovery of new small molecules that can target physiologically significant nucleic acid structures. A combined ligand and structure based virtual screening approach was utilized for the discovery of novel triplex binding small molecules. After screening several million small molecules and applying the *in silico* metrics, two small molecules were tested using a combination of biophysical techniques and were demonstrated to bind selectively to the triplex structure, as predicted by the *in silico* screen. These findings were a critical validation of the use of the *in silico* metrics for the discovery of new scaffolds of ligands that may have therapeutic value.

Finally, the last set of experiments describes the use of structure-based molecular docking approaches to identify new quadruplex binding small molecules. Over 6.6 million small molecules were docked into a quadruplex end-pasting site that we hypothesized to be structurally representative of the end-pasting site of the human telomeric G-quadruplex AGGG(TTAGGG)$_3$. A rank-by-rank consensus scoring function was used to re-rank the top hits from both Autodock and Surflex into a single top hit list. A single novel compound was discovered to bind to the human telomeric quadruplex by the *in silico* hypothesized end-pasting mechanism using a combination of biophysical techniques. This compound has a scaffold unlike any reported to date in the literature and represents a good lead for future derivatization experiments to further optimize the binding behavior of the compound.

The work presented has laid the foundation for future research investigating how small molecules interact with nucleic acids. We envision several areas in which this research can be well utilized. While the initial focus of this research was on the prediction of binding behavior of “pure” groove binders or classical intercalators, we believe this research can be extended to the field of “non-classical” intercalators, which are typified structurally by unfused polyaromatic ring systems and consist of a mixed-mode action with both groove binding and intercalation character. Compounds that bind by "non-classical intercalation" have garnered great interest recently because of their prevalence in the pharmaceutical industry. In fact, it was recently determined that as many as 26 out of 50 currently marketed drugs demonstrated surprising clastogenicity and these compounds were surprisingly mutagenic. This may be partly ascribed to the majority of these ligands having atypical, non-standard intercalator structures [223]. While there is software including DEREK, TOPKAT and MCASE that is used to estimate the toxicity of pharmaceutical compounds, these software have been largely unsuccessful at predicting the toxicity of non-classical intercalators [223]. It would be particularly valuable to extend the work performed here to the field of non-classical intercalators to determine if this binding mechanism can also be predicted by *in silico* based approaches.

The research here also has particular additional relevance in the field of quadruplex nucleic acids. A logical extension of our work is to develop metrics that can be used to discover small molecules that selectively discriminate between various quadruplex structures. Finding small molecules that can selectively bind to a specific G-quadruplex morphology has largely been elusive in the literature. The development of targeted metrics for predicting this behavior could be immensely powerful as virtual screening could potentially be used to identify new scaffolds of ligands that bind to a quadruplex of interest.

The research presented investigates the prediction of small molecule –nucleic acid interactions by *in silico* molecular docking and metric development. A unique facet of the work is the empirical validation of the *in silico* results using a number of spectroscopic, calorimetric and other techniques. This emphasizes that the work as described here has practical applications for the clinical discovery of new small molecules with therapeutic indications. There continues to be an increasing need to find new drugs to treat many types of disease. The research as outlined here provides a novel approach to discover small molecules to meet this need.

The combined approach of virtual screening and empirical validation of hits continues to identify many compounds with medicinal benefit. Our most recent work and our current focus is on investigating the potential anti-cancer properties of over 20 compounds that we have recently identified. Preliminary testing of these compounds has demonstrated that all of these compounds bind to the K^+ 22 mer human telomeric quadruplex and several compounds significantly stabilize the quadruplex, suggesting they may also inhibit telomerase and suppress tumor cell growth. The labs of Drs Trent and Chaires are now focused on rigorously testing these hits using a combination of spectroscopic, calorimetric and other biophysical techniques as well as multiple assays to measure cellular inhibition of telomerase (TRAP assay) and tumor cell proliferation (MTT assay). This is a practical application of our research and demonstrates that the integrated discovery and testing strategy we described here has led to the discovery of multiple novel small molecules with possible anti-cancer activity.

REFERENCES

1. Drews, J., *Drug discovery: a historical perspective.* Science, 2000. **287**(5460): p. 1960-4.
2. Detering, C. and G. Varani, *Validation of automated docking programs for docking and database screening against RNA drug targets.* J Med Chem, 2004. **47**(17): p. 4188-201.
3. Hurley, L.H., *DNA and its associated processes as targets for cancer therapy.* Nat Rev Cancer, 2002. **2**(3): p. 188-200.
4. Lauria, A., et al., *DNA minor groove binders: an overview on molecular modeling and QSAR approaches.* Curr Med Chem, 2007. **14**(20): p. 2136-60.
5. Reddy, B.S., S.M. Sondhi, and J.W. Lown, *Synthetic DNA minor groove-binding drugs.* Pharmacol Ther, 1999. **84**(1): p. 1-111.
6. Lane, A.N. and T.C. Jenkins, *Structures and Properties of Multi-stranded Nucleic Acids.* Current Organic Chemistry, 2001. **5**: p. 845-869.
7. Chaires, J.B., *Competition dialysis: an assay to measure the structural selectivity of drug-nucleic acid interactions.* Curr Med Chem Anticancer Agents, 2005. **5**(4): p. 339-52.
8. Hurley, L.H., et al., *G-quadruplexes as targets for drug design.* Pharmacol Ther, 2000. **85**(3): p. 141-58.
9. Weyermann, P. and P.B. Dervan, *Recognition of ten base pairs of DNA by head-to-head hairpin dimers.* J Am Chem Soc, 2002. **124**(24): p. 6872-8.
10. Ren, J. and J.B. Chaires, *Sequence and structural selectivity of nucleic acid binding ligands.* Biochemistry, 1999. **38**(49): p. 16067-75.
11. Kim, M.Y., et al., *The different biological effects of telomestatin and TMPyP4 can be attributed to their selectivity for interaction with intramolecular or intermolecular G-quadruplex structures.* Cancer Res, 2003. **63**(12): p. 3247-56.
12. Kinnings, S.L. and R.M. Jackson, *LigMatch: a multiple structure-based ligand matching method for 3D virtual screening.* J Chem Inf Model, 2009. **49**(9): p. 2056-66.
13. Jorgensen, W.L., *The many roles of computation in drug discovery.* Science, 2004. **303**(5665): p. 1813-8.
14. Ekins, S., J. Mestres, and B. Testa, *In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling.* Br J Pharmacol, 2007. **152**(1): p. 9-20.
15. Khadikar, P., et al., *QSAR studies on 1,2-dithiole-3-thiones: modeling of lipophilicity, quinone reductase specific activity, and production of growth hormone.* Bioorg Med Chem Lett, 2005. **15**(4): p. 1249-55.

16. de Groot, M.J. and S. Ekins, *Pharmacophore modeling of cytochromes P450.* Adv Drug Deliv Rev, 2002. **54**(3): p. 367-83.
17. Wang, S., et al., *The discovery of novel, structurally diverse protein kinase C agonists through computer 3D-database pharmacophore search. Molecular modeling studies.* J Med Chem, 1994. **37**(26): p. 4479-89.
18. Wielert-Badt, S., et al., *Probing the conformation of the sugar transport inhibitor phlorizin by 2D-NMR, molecular dynamics studies, and pharmacophore analysis.* J Med Chem, 2000. **43**(9): p. 1692-8.
19. Barreca, M.L., et al., *Pharmacophore modeling as an efficient tool in the discovery of novel noncompetitive AMPA receptor antagonists.* J Chem Inf Comput Sci, 2003. **43**(2): p. 651-5.
20. Aronov, A., *Applications of QSAR methods to ion channels.* 2006, Hoboken, NJ: John Wiley & Sons.
21. Tsuchida, K., et al., *Discovery of nonpeptidic small-molecule AP-1 inhibitors: lead hopping based on a three-dimensional pharmacophore model.* J Med Chem, 2006. **49**(1): p. 80-91.
22. Gopalakrishnan, B., et al., *A virtual screening approach for thymidine monophosphate kinase inhibitors as antitubercular agents based on docking and pharmacophore models.* J Chem Inf Model, 2005. **45**(4): p. 1101-8.
23. Steindl, T. and T. Langer, *Influenza virus neuraminidase inhibitors: generation and comparison of structure-based and common feature pharmacophore hypotheses and their application in virtual screening.* J Chem Inf Comput Sci, 2004. **44**(5): p. 1849-56.
24. Kapetanovic, I.M., *Computer-aided drug discovery and development (CADDD): In silico-chemico-biological approach.* Chem Biol Interact, 2008. **171**(2): p. 165-76.
25. Dror, O., et al., *Novel approach for efficient pharmacophore-based virtual screening: method and applications.* J Chem Inf Model, 2009. **49**(10): p. 2333-43.
26. Zoete, V., A. Grosdidier, and O. Michielin, *Docking, virtual high throughput screening and in silico fragment-based drug design.* J Cell Mol Med, 2009. **13**(2): p. 238-48.
27. Rester, U., *From virtuality to reality - Virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective.* Curr Opin Drug Discov Devel, 2008. **11**(4): p. 559-68.
28. Schapira, M., et al., *Rational discovery of novel nuclear hormone receptor antagonists.* Proc Natl Acad Sci U S A, 2000. **97**(3): p. 1008-13.
29. Tondi, D., et al., *Structure-based discovery and in-parallel optimization of novel competitive inhibitors of thymidylate synthase.* Chem Biol, 1999. **6**(5): p. 319-31.
30. Baxter, C.A., et al., *New approach to molecular docking and its application to virtual screening of chemical databases.* J Chem Inf Comput Sci, 2000. **40**(2): p. 254-62.
31. Lipinski, C.A., et al., *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.* Adv Drug Deliv Rev, 2001. **46**(1-3): p. 3-26.
32. Hurley, L.H., *Secondary DNA structures as molecular targets for cancer therapeutics.* Biochem Soc Trans, 2001. **29**(Pt 6): p. 692-6.

33. Chen, Q., R.H. Shafer, and I.D. Kuntz, *Structure-based discovery of ligands targeted to the RNA double helix.* Biochemistry, 1997. **36**(38): p. 11402-7.
34. Evans, D.A. and S. Neidle, *Virtual screening of DNA minor groove binders.* J Med Chem, 2006. **49**(14): p. 4232-8.
35. Moore, M.J., et al., *Synthesis of distamycin A polyamides targeting G-quadruplex DNA.* Org Biomol Chem, 2006. **4**(18): p. 3479-88.
36. Grootenhuis, P.D., et al., *Computerized selection of potential DNA binding compounds.* Anticancer Drug Des, 1990. **5**(3): p. 237-42.
37. Grootenhuis, P.D., et al., *Finding potential DNA-binding compounds by using molecular shape.* J Comput Aided Mol Des, 1994. **8**(6): p. 731-50.
38. Rohs, R., et al., *Molecular flexibility in ab initio drug docking to DNA: binding-site and binding-mode transitions in all-atom Monte Carlo simulations.* Nucleic Acids Res, 2005. **33**(22): p. 7048-57.
39. Chen, Q., I.D. Kuntz, and R.H. Shafer, *Spectroscopic recognition of guanine dimeric hairpin quadruplexes by a carbocyanine dye.* Proc Natl Acad Sci U S A, 1996. **93**(7): p. 2635-9.
40. Monchaud, D., et al., *Ligands playing musical chairs with G-quadruplex DNA: a rapid and simple displacement assay for identifying selective G-quadruplex binders.* Biochimie, 2008. **90**(8): p. 1207-23.
41. Gray, R.D. and J.B. Chaires, *Kinetics and mechanism of K+- and Na+-induced folding of models of human telomeric DNA into G-quadruplex structures.* Nucleic Acids Res, 2008. **36**(12): p. 4191-203.
42. Gray, R.D., J. Li, and J.B. Chaires, *Energetics and Kinetics of a Conformational Switch in G-Quadruplex DNA (dagger).* J Phys Chem B, 2009.
43. Lane, A.N., et al., *Stability and kinetics of G-quadruplex structures.* Nucleic Acids Res, 2008. **36**(17): p. 5482-515.
44. Chaires, J.B., et al., *Triplex selective 2-(2-naphthyl)quinoline compounds: origins of affinity and new design principles.* J Am Chem Soc, 2003. **125**(24): p. 7272-83.
45. Jemal, A., et al., *Cancer statistics, 2008.* CA Cancer J Clin, 2008. **58**(2): p. 71-96.
46. *Surveillance and Health Policy Research.* 2009, American Cancer Society, Inc.
47. *DevCan: Probability of Developing or Dying of Cancer Software.* 2008, Statistical Research and Applications Branch, NCI.
48. Abdelrahim, M., et al., *Angiogenesis: an update and potential drug approaches (review).* Int J Oncol, 2010. **36**(1): p. 5-18.
49. Bonner, J.A., et al., *Radiotherapy plus cetuximab for locoregionally advanced head and neck cancer: 5-year survival data from a phase 3 randomised trial, and relation between cetuximab-induced rash and survival.* Lancet Oncol, 2010. **11**(1): p. 21-8.
50. Monsuez, J.J., et al., *Cardiac side-effects of cancer chemotherapy.* Int J Cardiol, 2010.
51. Aiken, M.J., et al., *Doxorubicin-induced cardiac toxicity and cardiac rest gated blood pool imaging.* Clin Nucl Med, 2009. **34**(11): p. 762-7.
52. van den Berg, J.H., et al., *Future opportunities in preventing cisplatin induced ototoxicity.* Cancer Treat Rev, 2006. **32**(5): p. 390-7.
53. Neidle, S. and M.A. Read, *G-quadruplexes as therapeutic targets.* Biopolymers, 2000. **56**(3): p. 195-208.

54. Arola, A. and R. Vilar, *Stabilisation of G-quadruplex DNA by small molecules.* Curr Top Med Chem, 2008. **8**(15): p. 1405-15.
55. Huppert, J.L., *Four-stranded nucleic acids: structure, function and targeting of G-quadruplexes.* Chem Soc Rev, 2008. **37**(7): p. 1375-84.
56. Chin, J.Y., E.B. Schleifman, and P.M. Glazer, *Repair and recombination induced by triple helix DNA.* Front Biosci, 2007. **12**: p. 4288-97.
57. Kalota, A., S.E. Shetzline, and A.M. Gewirtz, *Progress in the development of nucleic acid therapeutics for cancer.* Cancer Biol Ther, 2004. **3**(1): p. 4-12.
58. Seidman, M.M. and P.M. Glazer, *The potential for gene repair via triple helix formation.* J Clin Invest, 2003. **112**(4): p. 487-94.
59. *Autodock.* [cited 2007 October 25]; Available from: http://autodock.scripps.edu/faqs-help/tutorial.
60. Jain, A.N., *Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine.* J Med Chem, 2003. **46**(4): p. 499-511.
61. Baraldi, P.G., et al., *DNA minor groove binders as potential antitumor and antimicrobial agents.* Med Res Rev, 2004. **24**(4): p. 475-528.
62. Stiborova, M., et al., *Molecular mechanisms of antineoplastic action of an anticancer drug ellipticine.* Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub, 2006. **150**(1): p. 13-23.
63. Holt, P.A., J.B. Chaires, and J.O. Trent, *Molecular docking of intercalators and groove-binders to nucleic acids using Autodock and Surflex.* J Chem Inf Model, 2008. **48**(8): p. 1602-15.
64. Holt, P.A., et al., *Discovery of novel triple helical DNA intercalators by an integrated virtual and actual screening platform.* Nucleic Acids Res, 2009. **37**(4): p. 1280-7.
65. Wang, J., P.A. Kollman, and I.D. Kuntz, *Flexible ligand docking: a multistep strategy approach.* Proteins, 1999. **36**(1): p. 1-19.
66. Spitzer, G.M., et al., *DNA minor groove pharmacophores describing sequence specific properties.* J Chem Inf Model, 2007. **47**(4): p. 1580-9.
67. Jain, A.N., *Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search.* J Comput Aided Mol Des, 2007. **21**(5): p. 281-306.
68. Bissler, J.J., *Triplex DNA and human disease.* Front Biosci, 2007. **12**: p. 4536-46.
69. Han, H. and L.H. Hurley, *G-quadruplex DNA: a potential target for anti-cancer drug design.* Trends Pharmacol Sci, 2000. **21**(4): p. 136-42.
70. Pfeffer, P. and H. Gohlke, *DrugScoreRNA--knowledge-based scoring function to predict RNA-ligand interactions.* J Chem Inf Model, 2007. **47**(5): p. 1868-76.
71. Tse, W.C. and D.L. Boger, *Sequence-selective DNA recognition: natural products and nature's lessons.* Chem Biol, 2004. **11**(12): p. 1607-17.
72. Yan, Z., et al., *Identification of an aminoacridine derivative that binds to RNA tetraloops.* J Med Chem, 2007. **50**(17): p. 4096-104.
73. Reha, D., et al., *Intercalators. 1. Nature of stacking interactions between intercalators (ethidium, daunomycin, ellipticine, and 4',6-diaminide-2-phenylindole) and DNA base pairs. Ab initio quantum chemical, density functional theory, and empirical potential study.* J Am Chem Soc, 2002. **124**(13): p. 3366-76.

74. Nelson, S.M., L.R. Ferguson, and W.A. Denny, *DNA and the chromosome - varied targets for chemotherapy.* Cell Chromosome, 2004. **3**(1): p. 2.
75. Chaires, J.B., *A thermodynamic signature for drug-DNA binding mode.* Arch Biochem Biophys, 2006. **453**(1): p. 26-31.
76. Wemmer, D.E., *Designed sequence-specific minor groove ligands.* Annu Rev Biophys Biomol Struct, 2000. **29**: p. 439-61.
77. Broyles, S.S., M. Kremer, and B.A. Knutson, *Antiviral activity of distamycin A against vaccinia virus is the result of inhibition of postreplicative mRNA synthesis.* J Virol, 2004. **78**(4): p. 2137-41.
78. Nelson, S.M., L.R. Ferguson, and W.A. Denny, *Non-covalent ligand/DNA interactions: minor groove binding agents.* Mutat Res, 2007. **623**(1-2): p. 24-40.
79. *Autodock*, v. 4, Editor. 2007, The Scripps Research Institute: La Jolla, CA.
80. *Surflex-Dock.* 2007, Tripos, Inc.
81. Park, H., J. Lee, and S. Lee, *Critical assessment of the automated AutoDock as a new docking tool for virtual screening.* Proteins, 2006. **65**(3): p. 549-54.
82. Moitessier, N., et al., *Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go.* British Journal of Pharmacology, 2007. **153**: p. S7-S26.
83. Morris, G.M., et al., *Automated Docking Using a Lamarkian Genetic Algorithm and an Empirical Binding Free Energy Function.* Journal of Computational Chemistry, 1998. **19**(14): p. 1639-1662.
84. Ruppert, J., W. Welch, and A.N. Jain, *Automatic identification and representation of protein binding sites for molecular docking.* Protein Sci, 1997. **6**(3): p. 524-33.
85. Huey, R., et al., *A semiempirical free energy force field with charge-based desolvation.* J Comput Chem, 2007. **28**(6): p. 1145-52.
86. Pham, T.A. and A.N. Jain, *Parameter estimation for scoring protein-ligand interactions using negative training data.* J Med Chem, 2006. **49**(20): p. 5856-68.
87. Halperin, I., et al., *Principles of docking: An overview of search algorithms and a guide to scoring functions.* Proteins, 2002. **47**(4): p. 409-43.
88. Acharya, K.R. and M.D. Lloyd, *The advantages and limitations of protein crystal structures.* Trends Pharmacol Sci, 2005. **26**(1): p. 10-4.
89. Warren, G.L., et al., *A critical assessment of docking programs and scoring functions.* J Med Chem, 2006. **49**(20): p. 5912-31.
90. *Maestro.* 2007, Schrodinger, LLC: San Diego, CA.
91. *Macromodel 7.0.* 1999, Schrodinger, Inc.
92. *SYBYL 7.3.* 2006, Tripos, Inc.
93. *Amber 8.* 2004, amber.scripps.edu.
94. *AutoDockTools.* 2007, The Scripps Research Institute.
95. Hetenyi, C. and D. van der Spoel, *Efficient docking of peptides to proteins without prior knowledge of the binding site.* Protein Sci, 2002. **11**(7): p. 1729-37.
96. Bailly, C., et al., *Molecular determinants for DNA minor groove recognition: design of a bis-guanidinium derivative of ethidium that is highly selective for AT-rich DNA sequences.* Biochemistry, 2005. **44**(6): p. 1941-52.
97. Jain, A.N., *Surflex Manual: Docking and Similarity.* 2007. p. 1-21.
98. Jain, A. Personal communication, 2008, University of California, San Francisco: San Francisco, CA.

99. Kellenberger, E., et al., *Comparative evaluation of eight docking tools for docking and virtual screening accuracy.* Proteins, 2004. **57**(2): p. 225-42.
100. *OpenBabel.* 2007, Free Software Foundation, Inc.: Boston, MA.
101. *iBabel.* 2007, Cambridge MedChem Consulting: Cambridgeshire, UK.
102. Kim, R. and J. Skolnick, *Assessment of programs for ligand binding affinity prediction.* J Comput Chem, 2008. **29**(8): p. 1316-31.
103. Gani, O.A., *Signposts of docking and scoring in drug design.* Chem Biol Drug Des, 2007. **70**(4): p. 360-5.
104. Teramoto, R. and H. Fukunishi, *Supervised consensus scoring for docking and virtual screening.* J Chem Inf Model, 2007. **47**(2): p. 526-34.
105. *Schrodinger.* 1999.
106. Bourdouxhe-Housiaux, C., et al., *Interaction of a DNA-threading netropsin-amsacrine combilexin with DNA and chromatin.* Biochemistry, 1996. **35**(14): p. 4251-64.
107. Grand, C.L., et al., *The cationic porphyrin TMPyP4 down-regulates c-MYC and human telomerase reverse transcriptase expression and inhibits tumor growth in vivo.* Mol Cancer Ther, 2002. **1**(8): p. 565-73.
108. Siddiqui-Jain, A., et al., *Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription.* Proc Natl Acad Sci U S A, 2002. **99**(18): p. 11593-8.
109. Monchaud, D. and M.P. Teulade-Fichou, *A hitchhiker's guide to G-quadruplex ligands.* Org Biomol Chem, 2008. **6**(4): p. 627-36.
110. Ihmels, H. and D. Otto, *Intercalation of Organic Dye Molecules into Double-Stranded DNA - General Principles and Recent Developments.* Top Curr Chem, 2005. **258**: p. 161-204.
111. Nguyen, B., et al., *Characterization of a novel DNA minor-groove complex.* Biophys J, 2004. **86**(2): p. 1028-41.
112. Sinha, R., et al., *The binding of DNA intercalating and non-intercalating compounds to A-form and protonated form of poly(rC).poly(rG): spectroscopic and viscometric study.* Bioorg Med Chem, 2006. **14**(3): p. 800-14.
113. Nguyen, B., et al., *Strong binding in the DNA minor groove by an aromatic diamidine with a shape that does not match the curvature of the groove.* J Am Chem Soc, 2002. **124**(46): p. 13680-1.
114. Ricci, C.G. and P.A. Netz, *Docking studies on DNA-ligand interactions: building and application of a protocol to identify the binding mode.* J Chem Inf Model, 2009. **49**(8): p. 1925-35.
115. *SYBYL 8.1.* 2008, Tripos, Inc.
116. Gyi, J.I., et al., *Comparison of the thermodynamic stabilities and solution conformations of DNA.RNA hybrids containing purine-rich and pyrimidine-rich strands with DNA and RNA duplexes.* Biochemistry, 1996. **35**(38): p. 12538-48.
117. Morris, G.M., et al., *Automated Docking Using a Lamarkian Genetic Algorithm and an Empirical Binding Free Energy Function.* J Comput Chem, 1998. **19**: p. 1639-1662.
118. Edwards, K.J., T.C. Jenkins, and S. Neidle, *Crystal structure of a pentamidine-oligonucleotide complex: implications for DNA-binding properties.* Biochemistry, 1992. **31**(31): p. 7104-9.

119. Coll, M., et al., *A bifurcated hydrogen-bonded conformation in the d(A.T) base pairs of the DNA dodecamer d(CGCAAATTTGCG) and its complex with distamycin.* Proc Natl Acad Sci U S A, 1987. **84**(23): p. 8385-9.

120. Shaw, N.N., H. Xi, and D.P. Arya, *Molecular recognition of a DNA:RNA hybrid: sub-nanomolar binding by a neomycin-methidium conjugate.* Bioorg Med Chem Lett, 2008. **18**(14): p. 4142-5.

121. Hendry, L.B., et al., *Small molecule intercalation with double stranded DNA: implications for normal gene regulation and for predicting the biological efficacy and genotoxicity of drugs and other chemicals.* Mutat Res, 2007. **623**(1-2): p. 53-71.

122. Tuite, E., et al., *Effects of minor and major groove-binding drugs and intercalators on the DNA association of minor groove-binding proteins RecA and deoxyribonuclease I detected by flow linear dichroism.* Eur J Biochem, 1997. **243**(1-2): p. 482-92.

123. Fox, K.R. and R.A.J. Darby, *Triple Helix-Specific Ligands.* Small molecule DNA and RNA binders; from synthesis to nucleic acid complexes. 2003: Wiley-VCH.

124. Ferguson, L.R. and W.A. Denny, *Genotoxicity of non-covalent interactions: DNA intercalators.* Mutat Res, 2007. **623**(1-2): p. 14-23.

125. Bernier, J.-L., J.-P. Henichart, and J.-P. Catteau, *Design, synthesis and DNA-binding capacity of a new peptidic bifunctional intercalating agent.* Biochem J, 1981. **199**: p. 479-484.

126. Debray, J., et al., *Synthesis and evaluation of fused bispyrimidinoacridines as novel pentacyclic analogues of quadruplex-binder BRACO-19.* Org Biomol Chem, 2009. **7**(24): p. 5219-28.

127. Howell, L.A., et al., *Synthesis and evaluation of 9-aminoacridines derived from benzyne click chemistry.* Bioorg Med Chem Lett, 2009. **19**(20): p. 5880-3.

128. Chilin, A., et al., *Synthesis and antitumor activity of novel amsacrine analogs: the critical role of the acridine moiety in determining their biological activity.* Bioorg Med Chem, 2009. **17**(2): p. 523-9.

129. Campbell, N.H., et al., *Structural basis of DNA quadruplex recognition by an acridine drug.* J Am Chem Soc, 2008. **130**(21): p. 6722-4.

130. Chaires, J.B., et al., *Structural selectivity of aromatic diamidines.* J Med Chem, 2004. **47**(23): p. 5729-42.

131. Wei, C., et al., *A spectroscopic study on the interactions of porphyrin with G-quadruplex DNAs.* Biochemistry, 2006. **45**(21): p. 6681-91.

132. Wei, C., et al., *Evidence for the binding mode of porphyrins to G-quadruplex DNA.* Phys Chem Chem Phys, 2009. **11**(20): p. 4025-32.

133. Bahr, M., et al., *Selective recognition of pyrimidine-pyrimidine DNA mismatches by distance-constrained macrocyclic bis-intercalators.* Nucleic Acids Res, 2008. **36**(15): p. 5000-12.

134. Jourdan, M., et al., *Threading bis-intercalation of a macrocyclic bisacridine at abasic sites in DNA: nuclear magnetic resonance and molecular modeling study.* Biochemistry, 1999. **38**(43): p. 14205-13.

135. Kerwin, S.M., et al., *Perylene diimide G-quadruplex DNA binding selectivity is mediated by ligand aggregation.* Bioorg Med Chem Lett, 2002. **12**(3): p. 447-50.

136. Baudoin, O., et al., *Stabilization of DNA Triple Helices by Crescent-Shaped Dibenzophenanthrolines.* Chem. Eur. J., 1998. **4**(8): p. 1504-1508.

137. Teulade-Fichou, M.P., et al., *Selective recognition of G-Quadruplex telomeric DNA by a bis(quinacridine) macrocycle.* J Am Chem Soc, 2003. **125**(16): p. 4732-40.

138. Fox, K.R., et al., *A molecular anchor for stabilizing triple-helical DNA.* Proc Natl Acad Sci U S A, 1995. **92**(17): p. 7887-91.

139. Keppler, M.D., et al., *Stabilization of DNA triple helices by a series of mono- and disubstituted amidoanthraquinones.* Eur J Biochem, 1999. **263**(3): p. 817-25.

140. Sun, D., et al., *Inhibition of human telomerase by a G-quadruplex-interactive compound.* J Med Chem, 1997. **40**(14): p. 2113-6.

141. Read, M.A., et al., *Molecular modeling studies on G-quadruplex complexes of telomerase inhibitors: structure-activity relationships.* J Med Chem, 1999. **42**(22): p. 4538-46.

142. Webb, M.R. and S.E. Ebeler, *Comparative analysis of topoisomerase IB inhibition and DNA intercalation by flavonoids and similar compounds: structural determinates of activity.* Biochem J, 2004. **384**(Pt 3): p. 527-41.

143. Perry, P.J., et al., *2,7-Disubstituted amidofluorenone derivatives as inhibitors of human telomerase.* J Med Chem, 1999. **42**(14): p. 2679-84.

144. Alcaro, S., et al., *Tetraplex DNA specific ligands based on the fluorenone-carboxamide scaffold.* Bioorg Med Chem Lett, 2007. **17**(9): p. 2509-14.

145. Sinha, R. and G.S. Kumar, *Interaction of isoquinoline alkaloids with an RNA triplex: structural and thermodynamic studies of berberine, palmatine, and coralyne binding to poly(U).poly(A)(*)poly(U).* J Phys Chem B, 2009. **113**(40): p. 13410-20.

146. Arora, A., et al., *Binding of berberine to human telomeric quadruplex - spectroscopic, calorimetric and molecular modeling studies.* FEBS J, 2008. **275**(15): p. 3971-83.

147. Bhadra, K., M. Maiti, and G.S. Kumar, *DNA-binding cytotoxic alkaloids: comparative study of the energetics of binding of berberine, palmatine, and coralyne.* DNA Cell Biol, 2008. **27**(12): p. 675-85.

148. Berge, T., et al., *Structural perturbations in DNA caused by bis-intercalation of ditercalinium visualised by atomic force microscopy.* Nucleic Acids Res, 2002. **30**(13): p. 2980-6.

149. Crow, S.D., et al., *DNA sequence recognition by the antitumor drug ditercalinium.* Biochemistry, 2002. **41**(27): p. 8672-82.

150. Paramasivan, S. and P.H. Bolton, *Mix and measure fluorescence screening for selective quadruplex binders.* Nucleic Acids Res, 2008. **36**(17): p. e106.

151. Colson, P., C. Bailly, and C. Houssier, *Electric linear dichroism as a new tool to study sequence preference in drug binding to DNA.* Biophys Chem, 1996. **58**: p. 125-140.

152. Fox, K.R., S.L. Higson, and J.E. Scott, *Methyl green and its analogues bind selectively to AT-rich regions of native DNA.* Eur J Histochem, 1992. **36**: p. 263-270.

153. Cheng, C.C., et al., *Design of antineoplastic agents on the basis of the "2-phenylnaphthalene-type" structural pattern. 2. Synthesis and biological activity*

studies of benzo]b]naphtho[2,3-d]furan-6,11-dione derivatives. J Med Chem, 1993. **36**(25): p. 4108-12.

154. Kluza, J., et al., *Induction of apoptosis by the plant alkaloid sampangine in human HL-60 leukemia cells is mediated by reactive oxygen species.* Eur J Pharmacol, 2005. **525**(1-3): p. 32-40.
155. Marshall, K.M. and L.R. Barrows, *Biological activities of pyridoacridines.* Nat Prod Rep, 2004. **21**(6): p. 731-51.
156. Seidman, M.M., et al., *The development of bioactive triple helix-forming oligonucleotides.* Ann N Y Acad Sci, 2005. **1058**: p. 119-27.
157. Buchini, S. and C.J. Leumann, *Recent improvements in antigene technology.* Curr Opin Chem Biol, 2003. **7**(6): p. 717-26.
158. Kim, H.G., et al., *Inhibition of transcription of the human c-myc protooncogene by intermolecular triplex.* Biochemistry, 1998. **37**(8): p. 2299-304.
159. Carbone, G.M., et al., *DNA binding and antigene activity of a daunomycin-conjugated triplex-forming oligonucleotide targeting the P2 promoter of the human c-myc gene.* Nucleic Acids Res, 2004. **32**(8): p. 2396-410.
160. Lacoste, J., J.C. Francois, and C. Helene, *Triple helix formation with purine-rich phosphorothioate-containing oligonucleotides covalently linked to an acridine derivative.* Nucleic Acids Res, 1997. **25**(10): p. 1991-8.
161. Marchand, C., et al., *Stabilization of triple helical DNA by a benzopyridoquinoxaline intercalator.* Biochemistry, 1996. **35**(15): p. 5022-32.
162. Moraru-Allen, A.A., et al., *Coralyne has a preference for intercalation between TA.T triples in intramolecular DNA triple helices.* Nucleic Acids Res, 1997. **25**(10): p. 1890-6.
163. Keppler, M., et al., *DNA triple helix stabilisation by a naphthylquinoline dimer.* FEBS Lett, 1999. **447**(2-3): p. 223-6.
164. Ren, J., C. Bailly, and J.B. Chaires, *NB-506, an indolocarbazole topoisomerase I inhibitor, binds preferentially to triplex DNA.* FEBS Lett, 2000. **470**(3): p. 355-9.
165. Strekowski, L., et al., *New triple-helix DNA stabilizing agents.* Bioorg Med Chem Lett, 2005. **15**(4): p. 1097-100.
166. Cassidy, S.A., L. Strekowski, and K.R. Fox, *DNA sequence specificity of a naphthylquinoline triple helix-binding ligand.* Nucleic Acids Res, 1996. **24**(21): p. 4133-8.
167. Keppler, M.D., et al., *DNA sequence specificity of triplex-binding ligands.* Eur J Biochem, 2003. **270**(24): p. 4982-92.
168. Strekowski, L., et al., *Bis-4-aminoquinolines: novel triple-helix DNA intercalators and antagonists of immunostimulatory CpG-oligodeoxynucleotides.* Bioorg Med Chem, 2003. **11**(6): p. 1079-85.
169. Jain, A.N., *Ligand-based structural hypotheses for virtual screening.* J Med Chem, 2004. **47**(4): p. 947-61.
170. Cleves, A.E. and A.N. Jain, *Robust ligand-based modeling of the biological targets of known drugs.* J Med Chem, 2006. **49**(10): p. 2921-38.
171. Cleves, A.E. and A.N. Jain, *Effects of inductive bias on computational evaluations of ligand-based modeling and on drug discovery.* J Comput Aided Mol Des, 2007.

172. Ragazzon, P. and J.B. Chaires, *Use of competition dialysis in the discovery of G-quadruplex selective ligands.* Methods, 2007. **43**(4): p. 313-23.
173. Ragazzon, P.A., N.C. Garbett, and J.B. Chaires, *Competition dialysis: a method for the study of structural selective nucleic acid binding.* Methods, 2007. **42**(2): p. 173-82.
174. Carrasco, C., et al., *Tight binding of the antitumor drug ditercalinium to quadruplex DNA.* Chembiochem, 2002. **3**(12): p. 1235-41.
175. Shi, X. and J.B. Chaires, *Sequence- and structural-selective nucleic acid binding revealed by the melting of mixtures.* Nucleic Acids Res, 2006. **34**(2): p. e14.
176. Murphy, P.M., et al., *Biarylpyrimidines: a new class of ligand for high-order DNA recognition.* Chem Commun (Camb), 2003(10): p. 1160-1.
177. Xing, F., et al., *Molecular recognition of nucleic acids: coralyne binds strongly to poly(A).* FEBS Lett, 2005. **579**(22): p. 5035-9.
178. Ren, J., et al., *Molecular recognition of a RNA:DNA hybrid structure.* J Am Chem Soc, 2001. **123**(27): p. 6742-3.
179. Arya, D.P., L. Xue, and P. Tennant, *Combining the best in triplex recognition: synthesis and nucleic acid binding of a BQQ-neomycin conjugate.* J Am Chem Soc, 2003. **125**(27): p. 8070-1.
180. Irwin, J.J. and B.K. Shoichet, *ZINC--a free database of commercially available compounds for virtual screening.* J Chem Inf Model, 2005. **45**(1): p. 177-82.
181. Trent, J.O., *Molecular modeling of drug-DNA complexes: an update.* Methods Enzymol, 2001. **340**: p. 290-326.
182. Chaires, J.B., ed. *Structural Selectivity of Drug-Nucleic Acid Interactions Probed by Competition Dialysis.* ed. Springer-Verlag. 2005, GMBH & Co.: Heidleberg.
183. Chaires, J.B., *A competition dialysis assay for the study of structure-selective ligand binding to nucleic acids.* Current Procotols in Nucleic Acid Chemistry, ed. S.L. Beaucage. Vol. Chapter 8, Unit 8.3. 2003.
184. Garbett, N.C., P.A. Ragazzon, and J.B. Chaires, *Circular dichroism to determine binding mode and affinity of ligand-DNA interactions.* Nat Protoc, 2007. **2**(12): p. 3166-72.
185. Wilson, W.D., et al., *Design of RNA Interactive Anti-HIV Agents: Unfused Aromatic Intercalators.* Med. Chem. Res., 1992. **2**: p. 102-110.
186. Eriksson, M. and B. Norden, *Linear and Circular Dichroism of drug-nuclecic acid complexes.* Methods in enzymology, 2001. **340**: p. 68-98.
187. Neidle, S., *Human telomeric G-quadruplex: the current status of telomeric G-quadruplexes as therapeutic targets in human cancer.* FEBS J, 2010. **277**(5): p. 1118-25.
188. Patel, D.J., A.T. Phan, and V. Kuryavyi, *Human telomere, oncogenic promoter and 5'-UTR G-quadruplexes: diverse higher order DNA and RNA targets for cancer therapeutics.* Nucleic Acids Res, 2007. **35**(22): p. 7429-55.
189. Qin, Y. and L.H. Hurley, *Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions.* Biochimie, 2008. **90**(8): p. 1149-71.
190. Cosconati, S., et al., *Tandem application of virtual screening and NMR experiments in the discovery of brand new DNA quadruplex groove binders.* J Am Chem Soc, 2009. **131**(45): p. 16336-7.

191. Neidle, S., *The structures of quadruplex nucleic acids and their drug complexes.* Curr Opin Struct Biol, 2009. **19**(3): p. 239-50.
192. Parkinson, G.N., M.P. Lee, and S. Neidle, *Crystal structure of parallel quadruplexes from human telomeric DNA.* Nature, 2002. **417**(6891): p. 876-80.
193. Phan, A.T., *Human telomeric G-quadruplex: structures of DNA and RNA sequences.* FEBS J, 2010. **277**(5): p. 1107-17.
194. Dai, J., M. Carver, and D. Yang, *Polymorphism of human telomeric quadruplex structures.* Biochimie, 2008. **90**(8): p. 1172-83.
195. Ambrus, A., et al., *Human telomeric sequence forms a hybrid-type intramolecular G-quadruplex structure with mixed parallel/antiparallel strands in potassium solution.* Nucleic Acids Res, 2006. **34**(9): p. 2723-35.
196. Dai, J., et al., *Structure of the Hybrid-2 type intramolecular human telomeric G-quadruplex in K+ solution: insights into structure polymorphism of the human telomeric sequence.* Nucleic Acids Res, 2007. **35**(15): p. 4927-40.
197. Dai, J., et al., *Structure of the intramolecular human telomeric G-quadruplex in potassium solution: a novel adenine triple formation.* Nucleic Acids Res, 2007. **35**(7): p. 2440-50.
198. Tuntiwechapikul, W., et al., *The influence of pH on the G-quadruplex binding selectivity of perylene derivatives.* Bioorg Med Chem Lett, 2006. **16**(15): p. 4120-6.
199. Shi, D.F., et al., *Quadruplex-interactive agents as telomerase inhibitors: synthesis of porphyrins and structure-activity relationship for the inhibition of telomerase.* J Med Chem, 2001. **44**(26): p. 4509-23.
200. Dixon, I.M., et al., *Porphyrin derivatives for telomere binding and telomerase inhibition.* Chembiochem, 2005. **6**(1): p. 123-32.
201. Parkinson, G.N., F. Cuenca, and S. Neidle, *Topology conservation and loop flexibility in quadruplex-drug recognition: crystal structures of inter- and intramolecular telomeric DNA quadruplex-drug complexes.* J Mol Biol, 2008. **381**(5): p. 1145-56.
202. Luedtke, N.W., *Targeting G-Quadruplex DNA with Small Molecules.* Chimia, 2009. **63**: p. 134-139.
203. Pan, J. and S. Zhang, *Interaction between cationic zinc porphyrin and lead ion induced telomeric guanine quadruplexes: evidence for end-stacking.* J Biol Inorg Chem, 2009. **14**(3): p. 401-407.
204. Dailey, M.M., et al., *Structure-based drug design: from nucleic acid to membrane protein targets.* Exp Mol Pathol, 2009. **86**(3): p. 141-50.
205. Goodsell, D.S., G.M. Morris, and A.J. Olson, *Automated docking of flexible ligands: applications of AutoDock.* J Mol Recognit, 1996. **9**(1): p. 1-5.
206. Teramoto, R. and H. Fukunishi, *Consensus scoring with feature selection for structure-based virtual screening.* J Chem Inf Model, 2008. **48**(2): p. 288-95.
207. Wang, R. and S. Wang, *How does consensus scoring work for virtual library screening? An idealized computer experiment.* J Chem Inf Comput Sci, 2001. **41**(5): p. 1422-6.
208. Hudson, J.S., S.C. Brooks, and D.E. Graves, *Interactions of actinomycin D with human telomeric G-quadruplex DNA.* Biochemistry, 2009. **48**(21): p. 4440-7.

209. Evans, S.E., et al., *End-stacking of copper cationic porphyrins on parallel-stranded guanine quadruplexes.* J Biol Inorg Chem, 2007. **12**(8): p. 1235-49.
210. Allen, D.L. and G.J. Pielak, *Baseline length and automated fitting of denaturation data.* Protein Sci, 1998. **7**(5): p. 1262-3.
211. Lumry, R. and R. Biltonen, *Validity of the "two-state" hypothesis for conformational transitions of proteins.* Biopolymers, 1966. **4**(8): p. 917-44.
212. Read, M., et al., *Structure-based design of selective and potent G quadruplex-mediated telomerase inhibitors.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 4844-9.
213. Oda, A., et al., *Comparison of consensus scoring strategies for evaluating computational models of protein-ligand complexes.* J Chem Inf Model, 2006. **46**(1): p. 380-91.
214. Suh, D. and J.B. Chaires, *Criteria for the mode of binding of DNA binding agents.* Bioorg Med Chem, 1995. **3**(6): p. 723-8.
215. Zhang, H., et al., *Conformational conversion of DNA G-quadruplex induced by a cationic porphyrin.* Spectrochim Acta A Mol Biomol Spectrosc, 2009. **74**(1): p. 243-7.
216. Monchaud, D., C. Allain, and M.P. Teulade-Fichou, *Development of a fluorescent intercalator displacement assay (G4-FID) for establishing quadruplex-DNA affinity and selectivity of putative ligands.* Bioorg Med Chem Lett, 2006. **16**(18): p. 4842-5.
217. Allain, C., D. Monchaud, and M.P. Teulade-Fichou, *FRET templated by G-quadruplex DNA: a specific ternary interaction using an original pair of donor/acceptor partners.* J Am Chem Soc, 2006. **128**(36): p. 11890-3.
218. Boger, D.L. and W.C. Tse, *Thiazole orange as the fluorescent intercalator in a high resolution fid assay for determining DNA binding affinity and sequence selectivity of small molecules.* Bioorg Med Chem, 2001. **9**(9): p. 2511-8.
219. Tse, W.C. and D.L. Boger, *A fluorescent intercalator displacement assay for establishing DNA binding selectivity and affinity.* Acc Chem Res, 2004. **37**(1): p. 61-9.
220. Rachwal, P.A. and K.R. Fox, *Quadruplex melting.* Methods, 2007. **43**(4): p. 291-301.
221. Guedin, A., P. Alberti, and J.L. Mergny, *Stability of intramolecular quadruplexes: sequence effects in the central loop.* Nucleic Acids Res, 2009. **37**(16): p. 5559-67.
222. Mashimo, T., et al., *Folding pathways of hybrid-1 and hybrid-2 G-quadruplex structures.* Nucleic Acids Symp Ser (Oxf), 2008(52): p. 409-10.
223. Snyder, R.D., *Assessment of atypical DNA intercalating agents in biological and in silico systems.* Mutat Res, 2007. **623**(1-2): p. 72-82.

CURRICULUM VITAE

PATRICK ANDREW HOLT

Email: pah13@yahoo.com

1418 Highland Avenue
Louisville, KY 40204
(c): 410.279.2303

EDUCATION

The University of Louisville, Louisville, KY
Doctor of Medicine (M.D.), Estimated Date of Completion: 2012
Doctor of Philosophy (Ph.D.), Estimated Date of Completion: 2010
Class Rank: 19 out of 147 (after second year of medical school)

The University of Louisville, Louisville, KY
Master of Science (M.S.), 2009.
Department: Biochemistry and Molecular Biology

The Johns Hopkins University, Baltimore, MD
Master of Science (M.S.), 2002.
Department: Biotechnology GPA: 4.0

The Johns Hopkins University, Baltimore, MD
Bachelor of Science (B.S.), 2000.
Department: Biomedical Engineering (Chemical Engineering concentration)

HONORS AND AWARDS

M.D./Ph.D. Program, The University of Louisville, (Aug 2005)

Director's Award Recipient, Process Biochemistry, MedImmune, Inc., (June 2002)

SCIENTIFIC EXPERIENCE

Ph.D. Candidate, Advisor: John Trent, Ph.D., Department of Biochemistry and Molecular Biology, University of Louisville, Louisville, KY (Jul 2007 – present)

M.D. Candidate, School of Medicine, University of Louisville, Louisville, KY (Aug 2005 – present)

Research Associate I - Associate Scientist I, Process Biochemistry, MedImmune Inc. (owned by AstraZeneca), Gaithersburg, MD (Mar 2001 – Jul 2005)

Chemical Engineering Research Assistant, Biological Science Group, Naval Surface Warfare Center Carderock Division, Bethesda, MD (June 2000 to March 2001)

Physiology Lab Research Assistant, School of Medicine, The Johns Hopkins University, Baltimore, MD (August 1999-May 2000)

Chemical Engineering Analyst, W.R. Grace, Baltimore, MD (summers 1998, 1999)

Toxicology Lab Research Assistant, School of Public Health and Hygiene, The Johns Hopkins University, Baltimore, MD (spring 1998)

PUBLICATIONS

Holt, P. A.; Ragazzon, P.; Strekowski, L.; Chaires, J. B.; Trent, J. O., Discovery of novel triple helical DNA intercalators by an integrated virtual and actual screening platform. *Nucleic Acids Res* **2009,** 37,
(4), 1280-7.

Holt, P. A.; Chaires, J. B.; Trent, J. O., Molecular docking of intercalators and groove-binders to nucleic acids using Autodock and Surflex. *J Chem Inf Model* **2008,** 48, (8), 1602-15.

Dailey, M. M.; Hait, C.; **Holt, P. A.**; Maguire, J. M.; Meier, J. B.; Miller, M. C.; Petraccone, L.; Trent, J. O., Structure-based drug design: from nucleic acid to membrane protein targets. *Exp Mol Pathol* **2009,** 86, (3), 141-50.

Cunningham, A., Qamar, S., Carrasquer, A., **Holt, P**., Maguire, J., Cunningham, S., Trent, J. Mammary Carcinogen-Protein Binding Potentials: Novel and Biologically Relevant Structure-Activity Relationship Model Descriptors. Manuscript Accepted for Publication in SAR and QSAR in Environmental Research, April 2010.

Wei Y., Chen J., Rosas G., Tompkins D., **Holt P**., Rao R. Phenotypic Screening Of Mutations In Pmr1, the Yeast Secretory Pathway Ca^{2+}/Mn^{2+}-ATPase, Reveals Residues Critical for Ion Selectivity and Transport. *J Biol Chem.* **2000,** 275, (31), 23927-32.

POSTER PRESENTATIONS

Patrick A. Holt, Jonathan B. Chaires, John O. Trent. "Small Molecule Targeting of G-Quadruplexes using Virtual and Actual Screening." 2009 *Second International Meeting on Quadruplex DNA*. Louisville, KY.

Patrick A. Holt, Patricia Ragazzon, Jonathan B. Chaires, John O. Trent. "A Combination of Virtual High-Throughput Screening and Competition Dialysis For Identifying Ligands to Target Nucleic Acids." 2008 *2nd Biochemistry & Molecular Biology Colloquium.* Starlight, IN.

AR Cunningham, S Qamar, CA Carrasquer, **PA Holt**, JM Maguire, SL Cunningham, N Malik and JO Trent. "Structure-Activity Relationship Analysis of Rat Mammary Carcinogens: Using Chemical-Protein Binding Potentials as Novel and Biologically Relevant Structure Descriptors." *2009 Brown Cancer Center Retreat.* Louisville, KY.

Milton J. Axley, **P. Andrew Holt**, Jose R Casas-Finet. "Light-Induced Oxidation of Proteins and Amino Acids in Solution." 2002 IBC Conference *The Impact of Post-Translational & Chemical Modifications on Protein Therapeutics*. San Diego, CA.

TEACHING EXPERIENCE

·**Medical Biochemistry**, School of Medicine, University of Louisville, Louisville, KY (Jan 2007 – May 2009)