Computational Chemistry in Biochemistry and Drug Design

Herbert J. Bernstein

School of Chemistry and Materials Science
Rochester Institute of Technology

hjbsch@rit.edu

Talk for CHMP 753

21 November 2017

Overview

Computational chemistry, especially computational physical chemistry, provides a set of tools that allow for highly accurate analysis of models of chemical systems. When biochemistry and rational drug design are focused on moderate-sized static systems and small dynamic systems, all the tools of computational physical chemistry can be applied and provide deep insights into the activity of biologically important molecules. As computers become faster, with more memory, more processors, bigger disks and faster networks, the definitions of "moderate-sized" and "small" cover larger and more complex systems, but for the foreseeable future the application of computational chemistry in biochemistry and rational drug design will require approximations and simplifications, resulting in less accurate, but still very useful, results.

Contents

Biomolecules

The subjects of biochemistry are biomolecules, the molecules that are the basis of life. Small molecules such as H_2O, $NaCl$ and sugars are just as important biomolecules as *RNAs*, *DNAs*, proteins, polysaccharides and other macromolecules, and they are much easier to model and analyze in detail mathematically or with a computer. In most cases, we do not even have an accurate three-dimensional atomic structures for most of the larger macromolecules. As of 19 November 2017, there were only 135,201 stuctures of biological macromolecules in the Protein Data Bank [Bernstein et al., 1977] [Berman et al., 2000], as contrasted with 891.8 million sequences in the European Nucleotide Archive [Toribio et al., 2017] as of 6 November 2017.
The most accurate biomolecule structures are obtained experimentally using x-ray crystallography, nuclear magnetic resonance, and cryo-electron microscopy.

The Protein Folding Problem I

- Since 1961, there has been an increasingly successful effort to obtain three-dimensional protein structures computationally from protein sequence data [Anfinsen et al., 1961] [Roche and McGuffin, 2016] [Finkelstein et al., 2017] [Krokhotin and Dokholyan, 2017].
- What Anfinsen established was that proteins, once unfolded, have a strong tendency to refold again to their original structure, *i.e.* that sequence essentially determines structure. Unfortunately MD simulations trying to reproduce nature *ab initio* often failed to provide the right answer or ran at rates that would have taken time greater than the life of the universe to produce any answer at all.

The Protein Folding Problem II

- What has rescued the effort in recent years has been the rapid growth of the Protein Data Bank to provide a large pool of known structures that could then be used to fold major portions of new sequences by homology to sequences with experimentally determined structures (template based modelling methods – TBM). *Ab initio modelling* is now called template-free modelling methods – FM.
- The risk we face from TBM, however, is that we are gradually populating the pool of experimentally determined structures needed for this process with presumed model structures, which may compromise future structure determinations.

The Protein Folding Problem III

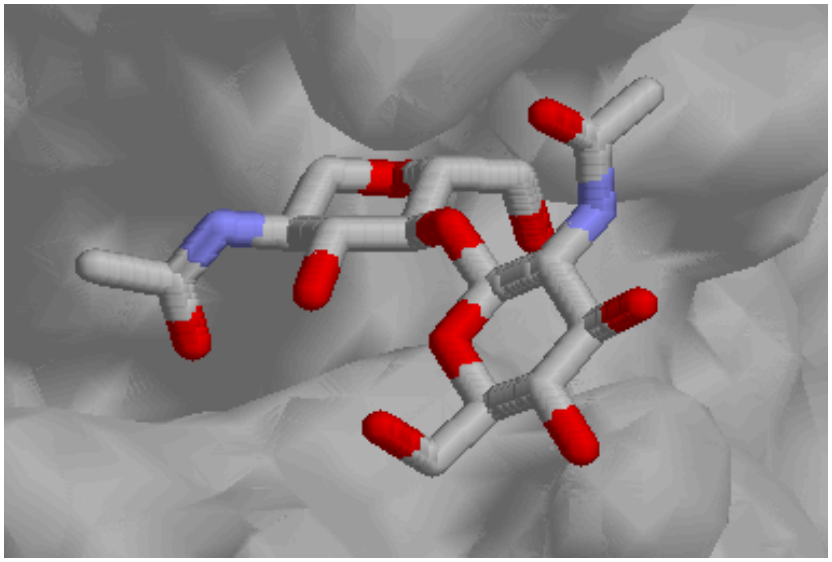
- There is a regular competition, Critical Assessment of Techniques for Structure Prediction, in which blind tests for both TBM and FM are conducted (see http://predictioncenter.org/casp12/zscores_final.cgi). The Rosetta server [Ovchinnikov et al., 2017] is currently the most effective tool.

Drug Discovery versus Drug Design

For most of history, new medications have been introduced to medical practice by hit-or-miss experimentation with new uses for existing drugs or by experimentation with minerals, plants, animal products and other substances found in nature, or one had to synthesize new substances. One had to extract or synthesize thousands of potential leads to slowly and patiently try *in vitro* (in the lab) or *in vivo* (in animals and human subjects) to discover useful activity and, hopefully, not injure or kill the subjects. This is the process of drug discovery. See
https://en.wikipedia.org/wiki/Drug_discovery

Ligands and Targets I

- The action of most drugs involves at least two molecules, the drug itself, and the "target," a molecule in a biological pathway the normal action of which the drug either inhibits or promotes. In many cases the drug is a ligand that binds to an **active site** on the target.



Usually the drug is a small molecule and the target is a macromolecule, most commonly a G-coupled-protein receptor (GCPR) or a kinase. This is a rendering of a portion of the surface of Protein Data Bank entry 5F19 [Lucido et al., 2016]

5F19 is "The Crystal Structure of Aspirin Acetylated Human Cyclooxygenase-2."

Ligands and Targets II

- As the three-dimensional structures of an increasing number of small molecules and macromolecules became known, it became increasingly feasible to move from screening *in vitro* or *in vivo* to a least preliminary screening *in silico*. This allowed rejection of the least promising leads. The combination of highly automated laboratory techniques, large databases of annotated chemical and biological data, and computational chemistry has resulted in high throughput screening. The molecules being screened may exist only as computer models, avoiding the need to actually synthesize those molecules unless the screening results look promising.

Differences in Representing Small Molecule Ligands and Macromolecule Targets I

- Much of rational drug design software is based on knowledge of three-dimensional atomic models of molecules. If you have accurate observations of element types and relative positions of atoms, you can make good estimates of bonding patterns and charges. With some techniques, you can even observe charge densities in addition to atomic positions.
- However, it becomes increasingly difficult to estimate the positions of individual atoms the larger a molecule gets. Therefore you may not have accurate observations of individual atom positions for the larger macromolecules. Instead what you are more likely to be able to observe are aggregations of atoms into groups or residues.

Differences in Representing Small Molecule Ligands and Macromolecule Targets II

- Therefore the primary representations of ligands are likely to be in terms of individual atoms (see the periodic table [Mendeleev, 1869] in https://en.wikipedia.org/wiki/Periodic_table and this interactive dynamic period table from http://www.ptable.com), while the primary representations of macromolecules are likely to be in terms of sequences of residues, either amino acids (see https://en.wikipedia.org/wiki/Amino_acid) or nuclei acids https://en.wikipedia.org/wiki/Nucleic_acid.
- This difference makes the representation of ligands simpler than the representation of macromolecules.
- This table of amino acids (from the RasMol Manual http://www.openrasmol.org/doc/helps us to understand how the most common residues in proteins will interact.

Differences in Representing Small Molecule Ligands and Macromolecule Targets III

Residues:	ala	arg	asn	asp	cys	glu	gln	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
Predefined Set																				
	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
acidic				*		*														
acyclic	*	*	*	*	*	*	*	*		*	*	*	*			*	*			*
aliphatic	*							*		*	*									*
aromatic									*					*				*	*	
basic		*							*			*								
buried	*				*					*	*		*	*				*		*
charged		*		*		*			*			*								
cyclic									*					*	*			*	*	
hydrophobic	*							*		*	*		*	*	*			*	*	*
large		*				*	*		*	*	*	*	*	*				*	*	
medium			*	*	*										*		*			*
negative				*		*														
neutral	*		*		*		*	*	*	*	*		*	*	*	*	*	*	*	*
polar		*	*	*	*	*	*		*			*				*	*			
positive		*							*			*								
small	*							*								*				
surface		*	*	*		*	*	*	*			*			*	*	*		*	

Rational Drug Design Processes I

- Designation of Targets
  - Molecules in a disease-related pathway for which changes in activity may impact the course of the disease.
  - The targets must be druggable.
  - May be identified by sequence or structual homology to known druggable targets.
  - Need to identify potential active sites.
- Designation of Drug Leads
  - Usually small molecule ligands.
  - Molecules complementary to identified active sites.
  - May have the structures of the targets to match against.
  - May have pharmacophore defined by other known ligands.
  - May be based on quantitative structure-activity relationship (QSAR) models.

Rational Drug Design Processes II

- If an accurate 3D target binding site structure is available, may **try virtual screening by docking models of ligands to the model of the binding site.**
- May screen to avoid binding to the wrong targets.
- Optimize lead characteristics to achieve druggability with acceptable toxicity.

A very popular virtual screening tool is Autodock Vina [Trott and Olson, 2010].

Autodock Vina I

"Molecular docking is a computational procedure that attempts to predict noncovalent binding of macromolecules or, more frequently, of a macromolecule (receptor) and a small molecule (ligand) efficiently, starting with their unbound structures, structures obtained from MD simulations, or homology modeling, etc. The goal is to predict the bound conformations and the binding affinity. "The prediction of binding of small molecules to proteins is of particular practical importance because it is used to screen virtual libraries of drug-like molecules in order to obtain leads for further drug development. ..."
– from the introduction to [Trott and Olson, 2010]

Accurate force fields are replaced by scoring functions used to approximate energies. The details are available in the paper https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041641/.

Autodock Vina II

Simplifying a simplification, it is fair to say Autodock Vina combines simple repulsive steric terms with hydrophobic terms and with hydrogen bonding.

Autodock Vina is a very popular internal engine for various scripts that run through large numbers of pair of ligands and targets, e.g. PyRx [Dallakyan and Olson, 2015].

Normally, we assume that there is a specific active site to which a specific ligand will dock in a specific orientation. There are, however, exceptions.

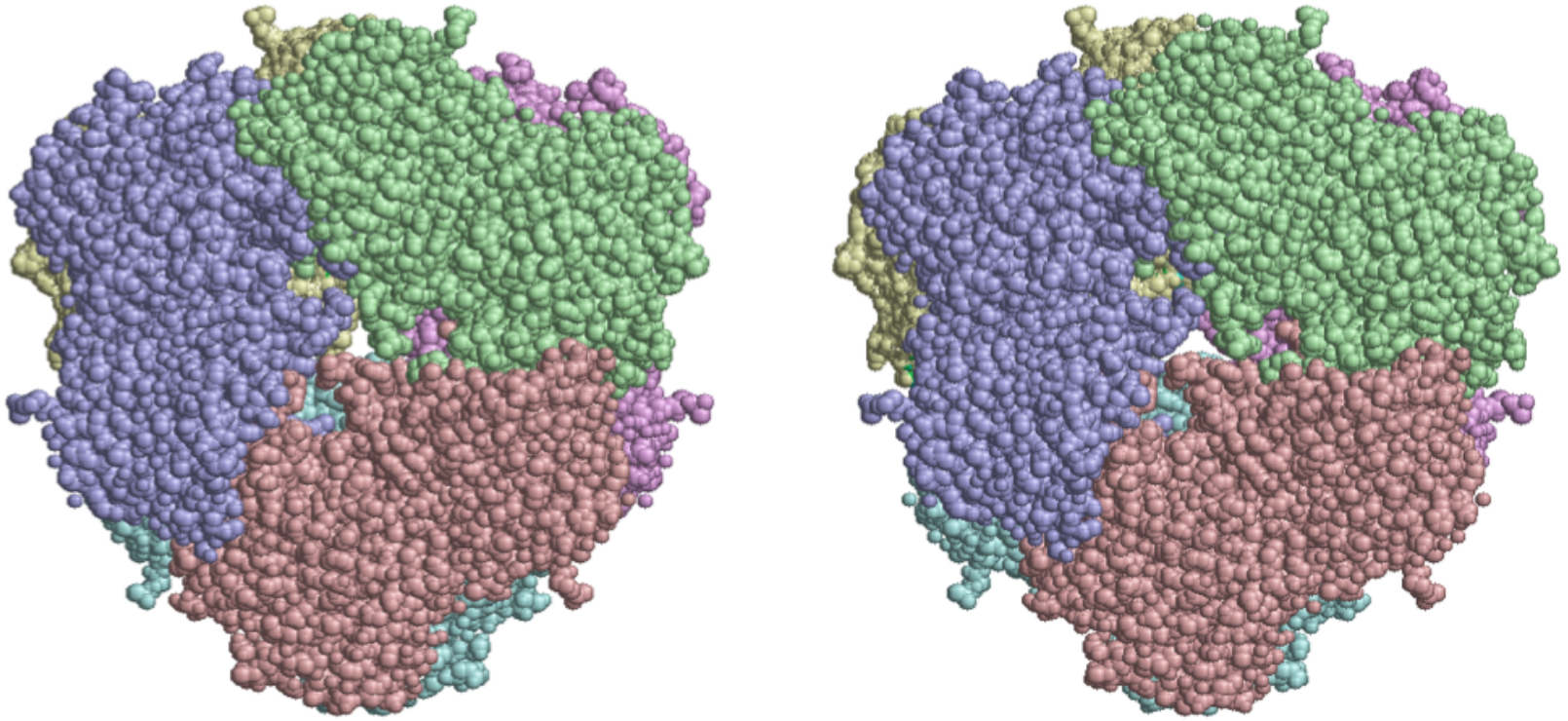
4GHB, A Promiscuous Docking Example I

"PDB entry 4GHB is a porin-like protein from *Bacteroides uniformis* ATCC 8492 with an unknown function [DOI:10.2210/pdb4ghb/pdb http://www.rcsb.org/pdb/files/4GHB.pdb]. We propose functions for this protein determined by use of *in silico* methods. One proposal is transport of NAP through cellular membrane. Porins are a class of proteins whose molecules can form channels large enough to allow the passage of small ions and molecules through cellular membranes. NAP corresponds to the enzyme cofactor nicotinamide adenine dinucleotide phosphate (NADP+), which serves as an electron carrier in redox reactions, alternating between oxidized (NADP+) and reduced (NADPH) forms. In silico mechanisms used are ProMOL and AutoDock. ProMOL is a plugin for the molecular visualization program, PyMOL, which uses catalytic site homology to predict the function of proteins with no known function ... . It defines catalytic sites using residue atom positions from known active sites. ProMOL

4GHB, A Promiscuous Docking Example II

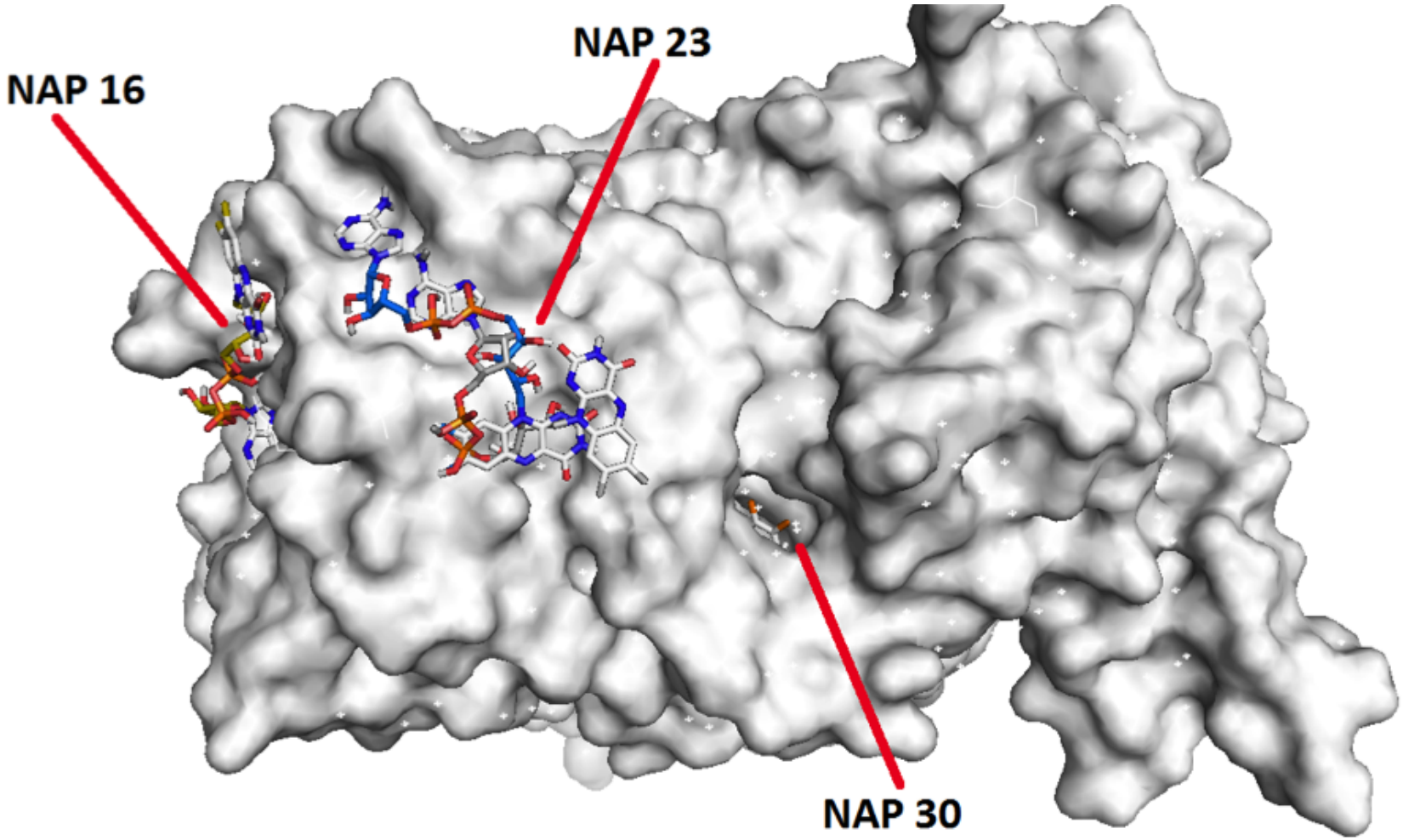
results for 4GHB suggested NAP as the ligand. To screen for ligand-receptor docking we used AutoDock
[Trott and Olson, 2010], which showed possible enzymatic activity at a particular active site [Cherry-Irby et al., 2015]. AutoDock revealed a complex binding pattern, with promiscuous binding at multiple sites in multiple orientations, rather than a single active site. This suggests a function of NAP transport."
– from the introduction to [Rosa et al., 2015]

4GHB, A Promiscuous Docking Example III



Pore formed from 6-mer of 4GHB

4GHB, A Promiscuous Docking Example IV



Promiscuous binding of NAP to 4GHB

Other Computational Issues in Biochemistry

- High dimension data visualization
- Extensive use of parallel computing
- Cluster analysis
- Post Office Problem
- Bellman's curse
- ...

References I

Anfinsen, C. B., Haber, E., Sela, M., and White, F. (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences*, 47(9):1309 – 1314.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235 – 242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer Jr., E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, 112(3):535 – 542.

Cherry-Irby, K., Craig, P., and Bernstein, H. (2015). *In silico* characterization and visualization of a protein of unknown function (4GHB) that is capable of enzymatic activity. In *American Crystallographic Association, Philadelphia, PA, July 2015.*

References II

Dallakyan, S. and Olson, A. J. (2015). Small-molecule library screening by docking with pyrx. *Chemical Biology: Methods and Protocols*, pages 243 – 250.

Finkelstein, A. V., Badretdin, A. J., Galzitskaya, O. V., Ivankov, D. N., Bogatyreva, N. S., and Garbuzynskiy, S. O. (2017). There and back again: Two views on the protein folding puzzle. *Physics of Life Reviews*, 21:56 – 71.

Krokhotin, A. and Dokholyan, N. V. (2017). Protein folding: Over half a century lasting quest: Comment on there and back again: Two views on the protein folding puzzle by alexei v. finkelstein et al. *Physics of Life Reviews*, 21:72 – 74.

Lucido, M. J., Orlando, B. J., Vecchio, A. J., and Malkowski, M. G. (2016). Crystal structure of aspirin-acetylated human cyclooxygenase-2: insight into the formation of products with reversed stereochemistry. *Biochemistry*, 55(8):1226 – 1238. PDB entry 5F19.

Mendeleev, D. (1869). The relation between the properties and atomic weights of the elements. *Journal of the Russian Chemical Society*, 1:60 – 77.

References III

Ovchinnikov, S., Park, H., Kim, D. E., DiMaio, F., and Baker, D. (2017). Protein structure prediction using rosetta in casp12. *Proteins: Structure, Function, and Bioinformatics*, pages 1 – 9.

Roche, D. B. and McGuffin, L. J. (2016). Toolbox for protein structure prediction. *Yeast Cytokinesis: Methods and Protocols*, pages 363 – 377.

Rosa, L., Craig, P., and Bernstein, H. (2015). *In silico* studies of the function of crystal structure of a porin-like protein(BACUNI_01323) from Bacteroides uniformis ATCC 8492 at 2.32 Åresolution. In *American Crystallographic Association, Philadelphia, PA, July 2015.*

Toribio, A. L., Alako, B., Amid, C., Cerdeño-Tarrága, A., Clarke, L., Cleland, I., Fairley, S., Gibson, R., Goodgame, N., ten Hoopen, P., et al. (2017). European nucleotide archive in 2016. *Nucleic acids research*, 45(D1):D32 – D36.

Trott, O. and Olson, A. J. (2010). Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455 – 461.

The End