Development of empirical molecular interaction models that incorporate hydrophobicity and hydropathy. The HINT paradigm

G.E. Kellogg* and D.J. Abraham

Institute for Structural Biology and Drug Discovery & Department of Medicinal Chemistry, School of Pharmacy, Virginia Commonwealth University, Richmond, VA 23298-0133, USA

While molecular mechanics calculates with impressive accuracy the small details of molecular structure, even for large proteins and other biological macromolecules, the gross features of these molecules are almost invariably poorly predicted. Most importantly molecular mechanics force fields do not include entropy. The HINT model, which we have been developing over the past several years, is based on the experimental measurement of log$P_{o/w}$, the partition coefficient for small molecule octanol/water solubilities. These partition coefficients represent a rare direct experimental measurement of non-covalent interaction that inherently includes entropy and solvation effects. This paper reviews the development of the HINT paradigm with illustrations from recent studies.

Introduction

One of the most puzzling aspects of modern computational chemistry is our continuing inability to predict the gross structure of biological macromolecules, especially proteins, from sequence [1]. This is puzzling because we can easily, with convincing accuracy, predict the microstructural features such as bond lengths, angles and conformation of individual sidechains. It is difficult to conceive of analogous situations in other fields of science; more often than not the "macro" is understood and manipulated much more readily than the "micro".

There are, of course, a multitude of reasons for this computational impotence. The complexity of biological systems is probably the most obvious. For a single protein we have to consider millions of potential contacts, thousands of degrees of freedom and likely hundreds of individual molecules if water is part of the model. The modeling of water with current methodology is difficult. Where should water be considered as bulk, which is conventionally dismissed by using a non-unity dielectric, and where should individual (ordered and disordered) water molecules be explicitly modeled? It is obvious that water is crucial; not only because of its role as a potent hydrogen bond mediator that supports protein structure, but also because significant entropy arises from displacement and/or movement of water in and out of the system [2]. The lack of adequate representation for entropy, despite the accuracy of the calculations for enthalpy, compromises our understanding of any real biomacromolecular structure or process. Complex dynamics simulations run over significant time frames would seem to be a reasonable response to these problems, but application continues to be expensive; and, even when the starting point of a dynamics simulation is the known crystallographic structure, after significant simulation time periods the final structure often loses much of the biomacromolecule's original secondary and tertiary structure.

For some time we have been interested in understanding and exploiting biological structure for a variety of purposes related to developing new disease treatments. The processes of ligand binding, protein-protein associations, protein-DNA associations and etc. are all fundamental to the design of new therapeutic agents. Our approach has been to develop a simple computational model that is based not on force fields, but on an actual *experimental* free energy measurement – the partition coefficient for octanol/water solubility, log$P_{o/w}$. This parameter has been used for several decades as a measure of lipid transport that can be often related to the *in vivo* biological activity of drug candidate molecules [3,4]. Log$P_{o/w}$ has also been shown in many cases to correlate with ligand binding measured *in vitro* [5,6]. At its simplest level log$P_{o/w}$ and its component fragment and atom terms reveal the type of interactions that the molecule/fragment/atom are able to make with another species. Log$P_{o/w}$ (or fragment/atom constants) less than zero is indicative of a polar (hydrophilic) species that would best interact with a polar environment containing hydrogen bond donors and/or acceptors. Positive log$P_{o/w}$ suggests a hydrophobic environment is appropriate. In addition, the distance from zero approximates the strength of potential interaction that could be made by the species if an appropriate partner is available. Formally charged atoms, fragments or molecules have significantly more negative log$P_{o/w}$ values than neutral (even polar) ones. This is in keeping with our expectation that such atoms or fragments are "richer" partners for hydrogen bonding.

Article available at http://analusis.edpsciences.org or http://dx.doi.org/10.1051/analusis:1999270019

Obviously, however, if there were no simple way to estimate the log$P_{o/w}$ value without actually measuring it for every molecule, this nascent model would have limited applicability. There are a growing variety of methods for prediction of log$P_{o/w}$ from structure and connectivity. Without reviewing a field already blessed with numerous reviews [7], these methods fall into two major types: atom-based and fragment-based. The difference is the definition of the minimum submolecular unit. In fragment methods, such as those by Rekker or Leo [8,9], the unit can be a fragment, which can be loosely identified with organic functional groups. These units are tied together by non-functional (carbon) atoms which have their own constants. The structural relationship between the fragments are encoded with "factors". The log$P_{o/w}$ for a molecule is calculated as the sum of all fragment constants and factors. In contrast, atom methods, such as by Ghose and Crippen and others [10-12], reparameterize organic chemistry into a fairly large number of specific atom types that encode environmental specifics. The log$P_{o/w}$ is the sum of all atom constants. These latter methods are easier to program, but can be less accurate with complex molecules involving multiple functional groups. However, as is often the case, we are interested in the relative differences in log$P_{o/w}$ among series of related molecules. All methods deal well with this type of problem.

The model we have been developing for biological interactions, HINT (for Hydropathic INTeractions), [13-17], uses an adaptation of the Hansch and Leo (CLOGP) [9,18] method for estimating log$P_{o/w}$. The atomistic hydrophobic constants that are the key parameter for the HINT model are obtained as described previously [13]. We term this value, *a*, the hydrophobic atom constant. It is derived from the Leo fragment constant *after* all of the Leo factors are applied. The sum of all hydrophobic atom constants is the log$P_{o/w}$ for the molecule. Then, a_i, the hydrophobic atom constant for atom *i*, encodes the energetics for interactions for *i*. We choose to modify these constants with the solvent-accessible surface area, *S*, for each atom. The primary reason is that the mantle or buried status of the atom, within its own fragment, directly affects its probability and strength of interaction. HINT calculates a score [a positive score represents a favorable interaction while a negative score represents an unfavorable interaction] for an atom-atom interaction as follows;

$$b_{ij} = a_i S_i a_j S_j f(r), \qquad (1)$$

where $f(r)$ is a function of the distance between atoms *i* and *j*. There are three possible scenarios: $a_i > 0$ & $a_j > 0$ (hydrophobic-hydrophobic), $a_{i/j} > 0$ & $a_{j/i} < 0$ (hydrophobic-polar) and $a_i < 0$ & $a_j < 0$ (polar-polar). This last case has three possible outcomes if we ascribe Lewis acid/base character to these polar atoms: acid-acid, acid-base or base-base. In the HINT model the hydrogen bond is a special case of acid-base determined by atom-atom distances, etc. Positive scores arise from hydrophobic-hydrophobic, acid-base or hydrogen bond; negative scores arise from hydrophobic-polar, acid-acid or base-base.

HINT is derived from a phenomenonological viewpoint. Thus we have pragmatically developed other aspects of the model, e.g., the distance relationship mediating hydropathic interactions through space. In our studies we most often use the simple exponential, $f(r) = e^{-r}$, to represent this behavior although the program is written to offer maximum flexibility to the user with a variety of both exponential and power functions. HINT also has an optional non-hydropathic-dependent term that is an implementation of the Lennard-Jones potential [19]. The purpose of this term is usually to flag and penalize van der Waals steric violations.

Scoring ligand binding with HINT

The original design of HINT was narrowly focused – to aid in the interpretation of a puzzling set of data for hemoglobin synthetic allosteric modulators. Two sets of compounds that differed only in the order -NHC(=O)- *vs*. -C(=O)NH- of an amide link between two substituted aromatic rings had substantially different measured biological activity as measured by p_{50} shifts of oxygen pressure [20]. X-ray crystallography of the hemoglobin-drug complexes [13] revealed that the carbonyl oxygen was directed in opposite directions in the two series, but provided no quantitative rationale for the biological activity differences. The HINT program identified, rationalized and quantified the important contacts effecting the drug binding [13]. The more active series was seen to have better access to the ammonium of Lys99α. Other important contacts were with Asn108β where the NH_2 donates into one of the drug aromatic rings, Arg141α which interacts with the drug carboxylate, and Phe36α defining the bottom of the hydrophobic pocket that the chlorinated phenyl ring occupies. The continuing development of this class of molecules has led to a new class of drug that is currently proceeding through clinical evaluation [21,22].

One calculated result from that study was the total "binding score" which is the sum of all individual atom-atom interactions, or $B = \Sigma\, b_{ij}$. Even in the early study that result showed promise as a method to rationally relate structure and interactions to biological activity (p_{50}). With the later availability of solution ligand binding data we reported correlation between the HINT scores and K_b [23]. To test this hypothesis further HINT was used to examine and score the output from the DOCK program [24]. Four ligand-receptor complexes of known structure were perturbed with DOCK and then the resulting complexes were scored with HINT. In three of the four cases, HINT identified the experimental binding mode. In the fourth case the experimental binding mode scored only slightly lower than a similar, energetically reasonable, orientation [25].

More recently, the HINT scoring paradigm has been incorporated into a novel methodology for data mining the National Cancer Institute repository of small molecules [26,27]. A three-dimensional stereoelectronic pharmacophore was developed from a quantitative structure-activity

relationship (QSAR) for inhibitors of human immunodeficiency virus type 1 (HIV-1) reverse transcriptase at the non-nucleoside binding site. The two-phase methodology used HINT score information in each phase. After the initial data base screen, which removed compounds with inappropriate $\log P_{o/w}$ or molecular volume, a 3D search using automated fitting protocols and scoring to select compounds for the second phase. The second, structural evaluation, phase examined conformational and complementarity of the putative active molecules with molecular modeling. This methodology was applied to a test database of about 10 000 compounds where 6.4% were known actives. All of the nineteen compounds passing the final modeling screen were known actives [27].

In other studies we have looked at the sequence specificity for doxorubicin intercalation into double helical DNA [28], developed theoretical binding sites and orientations for sugar binding to wheat germ agglutinin [29] and examined ligand binding to HIV-1 protease [30]. A detailed examination of the interactions between the molecular subunits of hemoglobin has shown that the HINT constants are scalable with ΔG for dimer-dimer assembly with a conversion factor of c.a. one kcal/mol being equivalent to 515 HINT score units [14].

HINT 3D property maps (including 3D QSAR)

It is a rather simple matter to convert an atom-based parameter into a "field" that can be contoured or otherwise visualized [31]. Once each atom of the target molecule has an intrinsic numerical property then one superimposes a 3D grid of test atoms over the molecule and surrounding space and calculates a set of "interactions" (for each atom of the molecule) at each grid point which are summed to create the field value at that point in space. In the case of HINT the atom's intrinsic property is $a_i S_i$. The test atoms are assumed to have $a_i = S_i = 1$; that is, they are small and hydrophobic. Thus the field value at grid point t is

$$A_t = \Sigma\, a_i\, S_i\, f(r_{it}). \qquad (2)$$

With HINT we generally use $f(r_{it}) = \exp(-r_{it})$ without the Lennard-Jones steric term for creating hydropathic property maps.

This approach can be used to nicely visualize the hydropathic profile of small molecules or macromolecules. Contour maps or slices show the expected trends of hydropathy; e.g., the interior of water soluble proteins is generally very hydrophobic reflecting the tendency of these proteins to expose their polar residues to solvent and thus create a hydrophobic core. For smaller molecules the contoured hydropathic maps reveal an easy to interpret three-dimensional pharmacophore of the molecule. The most significant application of these maps, however, is that they can be imported into 3D QSAR programs such as CoMFA [32] as hydropathic/hydrophobic fields to supplement the standard complement of steric and electrostatic field types. Since we first reported the HINT/CoMFA interface numerous studies have been performed using HINT [33-38] as well as other lipophilic field models [39,40] with 3D QSAR. Our experience is mixed. While the performance of the HINT field in CoMFA can be tuned with a variety of adjustable parameters, one can not *a priori* expect that the statistical measures of a 3D QSAR model will always be enhanced by the inclusion of the hydropathic field. However, in a few recent CoMFA studies, e.g., by Pajeva and Wiese [36], the HINT field was a significant component of the final model(s). The emerging general rule is that hydrophobic/hydropathic fields will be more applicable in cases where hydrophobic interactions are significant in the binding mode, or where membrane transport is encoded in the biological measure. On the other hand, the HINT and other hydrophobic fields uniquely report *chemical* information about the molecular system. And, as much as interpretability is important for drug or ligand design, these 3D QSAR fields produce the types of information that can guide chemists in producing new molecules.

Visualization of interactions

We have reported in several publications a new graphical display of interaction contours generated by HINT. This is accomplished by superimposing a grid over the region of interaction between two molecular species and considering each grid point to be an "observer" atom that monitors the through-space interactions in its vicinity. Basically the grid point is loaded with a distance mediated scalar value representing the summed score for all interactions between the interacting species. If the grid scalar value is positive that represents overall favorable interactions in that region of space. By sequentially calculating the interaction map for first "hydrophobic" where hydrophobic-hydrophobic (positive) and hydrophobic-polar (negative) interactions are encoded followed by "polar" where acid-base, hydrogen-bond (positive) and acid-acid, base-base (negative) interactions are encoded, we can display a clear and concise interaction map with visual data representing the types of interactions leading to the molecular association.

An illustration of such a map is set out as figure 1 (see page 39). This map shows the interactions between the ligand UK-129,485 (imidazodipyridodiazepine) and the non-nucleoside binding site of HIV-1 Reverse Transcriptase (Brookhaven Protein Data Bank entry code 1RVR) [26]. The types of interactions are coded as follows: hydrophobic-hydrophobic (green), hydrophobic-polar (purple), hydrogen bond and acid-base (blue), acid-acid and base-base (red). The ligand is displayed in ball and stick mode. From this map the interactions, both favorable and unfavorable, that effect its binding in the site are clear. The primary binding force is hydrophobic-hydrophobic interactions. Only a small proportion of hydrogen bonding/acid-base is present. Similarly, the unfavorable interactions are dominated by

hydrophobic-polar (in purple). Close examination of contour maps of this type can suggest, in a simple and easy to understand way, the types of chemical modifications to optimize the ligand. In this case the obvious goal would be to reduce the volume of the purple regions by adding the types of functional groups to the ligand that would complement the character of the enzyme binding site. We have calculated similar maps for the subunit-subunit interactions in hemoglobin tetramers [14] and from these we can propose site-directed mutations that could alter the energetics of the hemoglobin allosteric transition by stabilizing or destabilizing either the R or T endpoint states. HINT is one of the few, if not unique, computational programs that emphasizes the value of understanding and exploiting unfavorable as well as favorable interactions between species.

HINT future directions and availabilty

The future developments of the HINT model and algorithms are focusing on several areas. In most of our studies to date we have noted a divergence between molecular structures created or optimized using molecular mechanics, and structure scoring performed with HINT. This arises from the way that these two technologies treat interactions between hydrophobic atoms/groups and others. In general hydrophobic atoms will have a small positive charge, while many polar atoms (e.g., carbonyls, carboxylates, etc.) will have negative charges. As the primary algorithm for evaluating through-space interactions in most molecular mechanics force fields is the Coulombic potential, most hydrophobic-hydrophobic interactions are calculated as being energetically unfavorable and many hydrophobic-polar interactions are calculated as being energetically favorable. The HINT force field does not have that bias. One of the crucial long term goals of the HINT model is to evolve a full-featured HINT force field with a built-in minimizer. We believe that this could be of general utility in optimizing macromolecular crystal structures, and could be a useful tool in the quest for understanding protein folding. On another tack we are looking into new ways of simulating dielectric in interaction calculations. We are also importing the HINT model field into the HASL program of Arthur Doweyko [41,42]. HASL is a uniquely simple 3D QSAR algorithm that has been shown to give similar results to CoMFA, but with some advantages in terms of less complex statistics and the ability to spatially pinpoint regions of the molecule set responsible for the biological activity.

From the beginning our development of the HINT model has focused on usability issues. In fact, we started the small software company eduSoft, LC to license and package the HINT technology in commercial form. This has led us to hone the expertise for integration of our algorithms into the environments of popular commercial molecular modeling packages. We have developed HINT interfaces for Sybyl, insightII, and Chem-X over the past few years so that users of these modeling programs could add the features of HINT as easily as adding a new module to their system. To our knowledge we are unique in being able to provide our software to users with multiple interfaces and front ends. More information about eduSoft, LC and HINT can be found at *www.eslc.vabiotech.com*

Acknowledgements

We gratefully acknowledge the support of the HINT users who have continued to provide us with valuable feedback on our program. The HINT model and program was created with the advice and/or assistance of Simon Semus, Jeff Blaney, Peter Meiller and Rick Westkaemper. Our collaborators, Rick Gussio, Nagarajan Pattabiraman and Dan Zaharevitz of the National Cancer Institute, have supported the development of HINT by continuing to think of new ways to use it. The support and programming groups at Tripos, Inc., Molecular Simulations, Inc. and Chemical Design, Ltd. have helped with some aspects of the interface design and implementation. We acknowledge Virginia Commonwealth University and National Institutes of Health (HLBI grant RO-1-32793) for financial support.

References

1. Peitsch, M. C. Comparative Protein Modeling. in Practical Application of Computer-Aided Drug Design; Charifson P. S., Ed; Marcel Dekker, Inc., New York, 1997; 227-241.
2. Tanford, C. The Hydrophobic Effect: Formation of Micelles and Biological Membranes; John Wiley and Sons, Inc., New York, 1980.
3. Lien, E. J.; Gao, H. *Pharm. Res.* **1995**, *12*, 583-587.
4. Debnath, A. K.; Debnath, G.; Shusterman, A. J.; Hansch, C. *Environ. Mol. Mutagen.* **1992**, *19*, 37-52.
5. Kim, K. H.; Martin, Y. C.; Norris, B.; Young, P. R.; Carter, G. W.; Hasiv, F.; Walters, R.L. *J. Pharm. Sci.* **1990**, *79*, 609-613.
6. Mor, M.; Bordi, F.; Silva, C.; Rivara, S.; Crivori, P.; Plazzi, P. V.; Ballabeni, V; Caretta, A.; Barocelli, E. *J. Med. Chem.* **1997**, *40*, 2571-2578.
7. Mannhold, R.; Dross, K. *Quant. Struct-Act Relat.* **1996**, *15*, 403-408.
8. Rekker, R. F. The Hydrophobic Fragmental Constant, Its Derivation and Application: A Means of Characterizing Membrane Systems. Elsevier Scientific Pub., Amsterdam, 1977.
9. Hansch, C.; Leo, A. J. Substituent Constants for Correlation Analysis in Chemistry and Biology, John Wiley and Sons, Inc. New York, 1979.
10. Ghose, A. K.; Crippen, G. M. *J. Comp. Chem.* **1986**, *7*, 565-577.
11. Sasaki, Y.; Kubodera, H.; Matuszaki, T.; Umeyama, H. *J. Pharmacobiodyn.* **1991**, *14*, 207-214.
12. Gombar, V. K.; Enslein, K. *J. Comp. Inf. Comput. Sci.* **1996**, *36*, 1127-1134.
13. Kellogg, G. E.; Wireko, F. C.; Abraham, D. J. *J. Med. Chem.* **1991**, *34*, 758-767.
14. Abraham, D. J.; Kellogg, G. E.; Holt, J. M.; Ackers, G. K. *J. Mol. Bio.* **1997**, *272*, 613-632.

15. Kellogg, G. E.; Abraham, D. J. *J. Mol. Graph.* **1992**, *10*, 212-217
16. Abraham, D. J.; Kellogg, G. E. Hydrophobic Fields. In: 3D QSAR in Drug Design, Theory, Methods and Applications; Kubinyi H. Ed., ESCOM Sciences Publishers BV, Leiden, 1993; pp 506-522.
17. Abraham, D. J.; Kellogg, G. E. *J. Comp. Aid. Molec. Des.* **1994**, *8*, 41-49.
18. Abraham, D. J.; Leo, A. J. *Proteins* **1987**, *2*, 130-152.
19. Levitt, M. *J. Mol. Biol.* **1983**, *168*, 595-620.
20. Randad, R. S.; Maharan, M. A.; Mehanna, A. S.; Abraham, D. J. *J. Med. Chem.* **1991**, *34*, 752-757.
21. Watson, J. C.; Doppenberg, E. M. R.; Bullock, M. R.; Zauner, A.; Rice, M. R.; Abraham, D.; Young, H. F. *Stroke* **1997**, *28*, 1625-1631.
22. (a) Kunert, M. P.; Liard, J. F.; Abraham, D. J.; Lombard, J. H. *Microvascular Res.* **1996**, *52*, 58-68; (b) Mejia M. A.; Gerstenblith G.; Chacko V. P.; diPaula, A. F.; Becker, L. C.; Weiss, R. G. *Circulation* **1996**, *94* (Suppl I) 607.
23. Kellogg, G. E.; Joshi, G. S.; Abraham, D. J. *Med. Chem. Res.* **1992**, *1*, 444-453.
24. Kuntz, I. D.; Meng, E. C.; Shoichet, B. K. *Acc. Chem. Res.* **1994**, *27*, 117-125.
25. Meng, E. C.; Kuntz, I. D.; Abraham, D. J.; Kellogg, G. E. *J. Comp. Aid. Molec. Des.* **1994**, *8*, 299-306.
26. Gussio, R.; Pattabiraman, N.; Zaharevitz, D. W.; Kellogg, G. E.; Topol, I.; Rice, W. G.; Schaeffer C. A.; Erickson, J. W.; Burt, S. K. *J. Med. Chem.* **1996**, *39*, 1645-1650.
27. Gussio, R.; Pattabiraman, N.; Kellogg, G. E.; Zaharevitz, D. W. *Methods* **1998**, *14*, 255-263.
28. Kellogg, G. E.; Scarsdale, J. N.; Fornari, F. A. Jr. *Nucl. Acids Res.*, submitted.
29. Wright, C. S.; Kellogg, G. E. *J. Prot. Sci.* **1996**, *5*, 1466-1476.
30. Wei, D. T.; Kellogg, G. E.; Meadows, J. C. *Med. Chem. Res.* **1997**, *7*, 259-270.
31. Waller, C. L.; Kellogg, G. E. *NetSci*. **1996**, *2* (www.netsci.org/Science/Compchem/feature10.html).
32. Cramer, R. D., III; Patterson, D. E., Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959-5967.
33. Kellogg G. E.; Semus, S. F.; Abraham, D. J. *J. Comp. Aid. Molec. Des.* **1991**, *5*, 545-552.
34. Nayak, V. R.; Kellogg, G. E. *Med. Chem. Res.* **1994**, ***3***, 491-502.
35. Kellogg, G. E. *Med. Chem. Res.* **1997**, *7*, 417-427.
36. Pajeva, I.; Wiese, M. *J. Med. Chem.* **1998**, *41*, 1815-1826.
37. Welch, W.; Ahmad, S.; Gerzon, K.; Humerickhouse, R. A.; Besch, H. R., Jr.; Ruest, L.; Deslongchamps, P.; Sutko, J. L. *Biochemistry* **1994**, *33*, 6074-6085.
38. Oprea, T. I.; Waller, C. L.; Marshall, G. R. *Drug Design Discovery* **1994**, *12*, 29-51.
39. Gaillard, P.; Carrupt, P. -A.; Testa, B.; Schambel, P. *J. Med. Chem.* **1996**, *39*, 126-134.
40. Norinder, U. *J. Comp. Aid. Molec. Des.* **1991**, *4*, 381-389.
41. Doweyko, A. M. *J. Med. Chem.* **1988**, *31*, 1396.
42. Kaminski, J. J.; Doweyko, A. M. *J. Med. Chem.* **1997**, *40*, 427-436.