DRUG DISCOVERY TODAY BIOSILICO

Predicting aqueous solubility from structure

John S. Delaney

The aqueous solubility of a drug is one of the key physical properties that affect both its ADME profile and 'screenability' in HTS. This review critically surveys a range of methods that can be used to predict the solubility of a compound in water and presents some of the main issues that affect the applicability of different techniques. As ever, there are trade-offs to be made between the speed, accuracy and transparency of methods, but current programs can provide estimates to well within an order of magnitude in favourable cases. The need for new ways to predict solubility in more challenging systems (e.g. solvents such as DMSO and charged solutes) is discussed.

Water is integral to the structure and function of all living material. The human body is approximately 60% water by mass [1] and biology might be reasonably characterized as the 'damp science'. For the pharmaceutical industry, the behaviour of a drug in water governs many uptake, movement and elimination issues within the body (e.g. oral absorption and movement through blood), which affect the latter stages of the drug development process, as well as simple 'screenability' issues at the earliest (high-throughput screening) stages. Some of problems associated with combinatorial libraries identified by Lipinski [2] are closely related to their poor solubility. Poor solubility also affects the development of other commercially important compounds such as agrochemicals [3].

Given that the solubility of a compound can be of considerable interest, how do we go about obtaining it? The most obvious answer is to measure it directly. The traditional equilibrium approach requires a fair amount of sample (1–2 mg) and is time consuming (tens of hours to do properly). The time issue can thwart attempts to measure the solubility of compounds with limited stability in water: will the sample hydrolyse before equilibrium is reached? Fast, miniaturized methods such as nephelometry [4] can provide a kinetic solubility measurement with little starting material, but they require a reliable DMSO stock solution and multiple repeats to achieve accuracy. Another approach is to try to estimate solubility from more easily obtained measurements. The classic way to achieve this is to combine the log P (log of the partition coefficient of the compound between water and octanol) and melting point using the 'General Solubility Equation' (GSE [5]). The log P of a compound can be measured quite quickly using HPLC based methods [6] and melting point determination is a staple of any undergraduate chemistry course (heat the sample, watch it melt, note the temperature). Compounds with very high melting points (the sample decomposes before melting) or very low/high log Ps can cause problems, and it should be noted that the accuracy of the GSE applied to drug-sized compounds has been questioned [7].

All of the above presupposes that you have a sample of the compound – but if you don't, what then? High throughput screening has driven the rise in prominence of the solubilized collection – one where the only sample sits in DMSO solution at a questionable [8] concentration. There are more compounds that we would like to assess than we have the capacity to make, even with the rise of combinatorial chemistry. Molecules are cherry-picked from compound

John S. Delaney
Syngenta,
Jealott's Hill International
Research Centre,
Bracknell, Berkshire
UK, RG42 6EY
e-mail:
john.delaney@syngenta.com

1359-6446/04/$ – see front matter ©2005 Elsevier Ltd. All rights reserved. PII: S1359-6446(04)03365-3

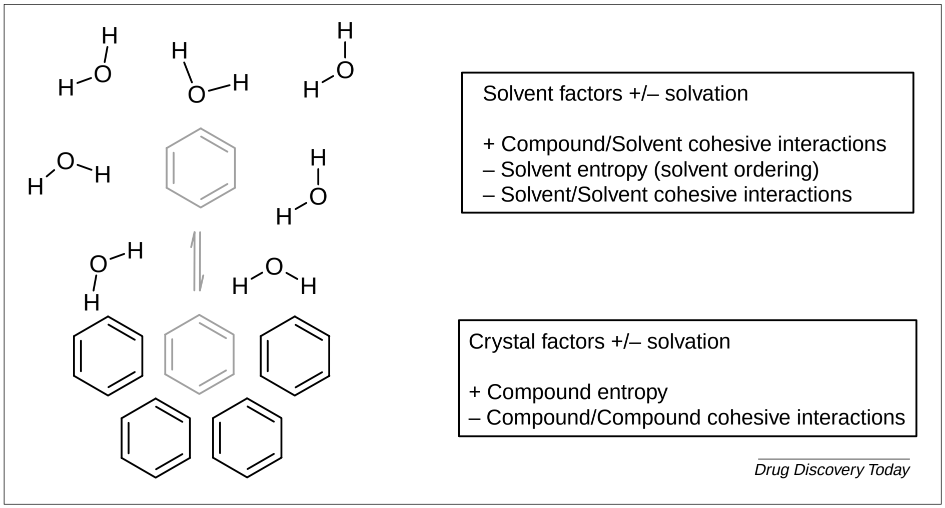
Solvent factors +/– solvation
+ Compound/Solvent cohesive interactions
– Solvent entropy (solvent ordering)
– Solvent/Solvent cohesive interactions
Crystal factors +/– solvation
+ Compound entropy
– Compound/Compound cohesive interactions
Drug Discovery Today

FIGURE 1

The main factors influencing the solvation of a crystalline compound. The + symbol refers to factors that favour the movement of the compound from the crystal lattice (bottom) into the solvent (top), the – symbol refers to factors that favour partition of compound into the lattice.

vendor databases and their suitability judged purely by their reported structure. All of this drives a demand for accurate ways to predict solubility directly from structure. In a recent review, Clark and Grootenhuis [9] noted a rise in the number of published methods for predicting solubility, a sign of the increased interest in the area. The remainder of this review will concern itself with solubility estimation in the absence of sample and attempt to survey the major computational methods for doing this.

Theory and methods

The dissolution of a drug in water is controlled by two kinds of interaction. First, we must consider how strongly the molecule associates with the solvent. Compounds containing large numbers of polar groups such as sugars might be expected to find more favourable interactions with water than unadorned hydrocarbons, and this is often translated into greater solubility. The other effect that controls solubility is the affinity of the solute for itself, or how tightly bound the compound is to its own crystal lattice. If the crystal's intermolecular interactions are strong, more energy will be required to wrench molecules out of it, leading to lower solubility. The balance between compound/solvent and compound/compound effects is illustrated schematically in Figure 1 and described more fully below.

Compound/solvent effects

The interactions between water and drug have been extensively studied and there are several ways to estimate this part of the problem. The simplest is to use log P, as Yalkowsky [10] has shown that log P can reasonably stand in for the activity coefficient in the overall solubility equation and this provides an estimate of the strength of the interaction of the compound with water. Most common log P estimation programs are fragment based and empirical, and there are several commercially available packages such as CLOGP (Daylight Chemical Information Systems) and ACD/LogD (Advanced Chemistry Development, Inc). CLOGP, in particular, has been used extensively in the pharmaceutical industry for many years and its strengths and weaknesses are well understood.

Molecular simulation offers another route to assessing the energetics of a compound in water. The model of the individual molecule in these simulations is relatively crude, but this is compensated for by simulating large ensembles of particles, allowing a statistical thermodynamic approach to be adopted. The work of Jorgensen and Duffy [11] exemplifies this approach, using Monte Carlo simulation (a technique that relies on randomly moving elements around to find low energy states and generate probability distributions – named after the famous casino in Monaco) with solute embedded in a bath of rigid water molecules to derive cohesive properties that can be used to predict solubility. An attractive feature of this method is that it takes some of the entropic contributions to the free energy into account directly. The main drawback of this approach is the computational load imposed by performing relatively complex simulations for each different solute. Jorgensen and Duffy have addressed this by using simulations as a way of deriving properties that can be used in a fast QSPR (quantitative structure property relationship) model (QikProp, Schrodinger, L.L.C.). QSPR models relate a set of molecular descriptor to a physical property of interest, in this case solubility. A completely different approach to simulation is offered by cellular automata [12], where physics is modelled as an extension of Conway's 'Game of Life' [13] – solvent and solute are represented by cells on a grid, their movement governed by their immediate neighbours and a set of transition rules. The 'game' is played out as a long series of steps and occupancy patterns of the cells change at each step. This type of simulation offers intriguing insights into the dissolution process (e.g. the formation of mobile cavities within the solid solute), but is probably not as useful as Monte Carlo for quantitative work.

An alternative to using a large number of crude particles is to use a single solute molecule modelled in more detail. This falls under the purview of quantum mechanical (QM) techniques with their sophisticated representation of molecules. A QM calculation attempts to describe a molecule's full quantum mechanical wave function, which is about as close to a true representation of a compound as current physics allows. Standard QM calculations are performed on compounds in the gas phase so there needs to be some way of accounting for the effects of solvent. The solvent both polarizes the molecule and is itself polarized by the solute, which considerably complicates the calculations. The Cramer-Truhlar [14] approach performs the QM calculation assuming that the compound is embedded in a continuous dielectric, which allows the polarization of the compound to be more accurately modelled. Klamt and co-workers [15] have produced the QM-based COSMO-RS (COSMOlogic GmbH and Co. KG) method, which goes further by embedding both solute and solvent in a perfect conductor to calculate their polarization charge densities. Integrating over the two surfaces (solute and solvent) allows the method to calculate the chemical potential of the solute in solvent, leading to an estimation of its solubility. As with Monte Carlo simulation, quantum mechanical methods are inherently slow and are not suitable for profiling large numbers of compounds.

Compound/compound effects

The interactions within the crystal lattice have received less attention in the solubility literature and this is somewhat understandable given the nature of the problem. Predicting crystal forms de-novo is a classically difficult computational problem, especially for molecules that exhibit any conformational flexibility. Given that the form is unpredictable, it follows that accurate estimates of the enthalpy of melting are practically impossible. Qualitative factors that contribute, such as numbers of hydrogen bond donor/acceptor groups, simply do not capture the full complexity of the state, although group contribution approaches might have something to offer [16]. In contrast to the water–drug situation, the entropy of the process might be more easily estimated. Yalkowsky and Dannenfelser [17] produced an estimate of ΔS_m (entropy of melting) based on the number of rotatable bonds and the symmetry of a compound. This type of approach is used in COSMO-RS and QikProp to factor in the lattice contribution to solvation. The relative importance of this term increases with larger molecules and it could be a significant component of drug solubility.

The strength of the crystal lattice can be represented by the melting point of the compound. Most of the work done on estimating melting point has used small, symmetric molecules or simple alkanes [18]. The results do not seem readily translatable to drug-like compounds. If the enthalpy and entropy of melting could be estimated accurately, then we might have a decent way of estimating the melting point of a compound. Work using this approach has so far produced a method with an accuracy of around +/–35°C [19], which equates to an error in GSE solubility estimate of around 0.3 log units. A recent empirical method [20] has achieved similar levels of accuracy. While these methods show promise, they have some way to go before an in-silico version of the GSE can compete with other computational solubility methods. In practice, most of the variance in the solubility of drug-like molecules comes from the water-solute term, which means that useable estimates can be obtained by essentially guessing the melting point! One can use a median melting point value (125°C is commonly used at Syngenta) in the GSE with a calculated log P to give an approximate solubility value. This works because the variance due to melting point in the GSE is around half a log unit, set against a variance due to log P of over 2 log units (figures based on an analysis of ten thousand compounds from the Syngenta corporate collection). The strong dependence of solubility on properties such as log P has formed the basis of several methods where calculated log P is augmented with additional terms [21–23]. These techniques are essentially empirical, although there have been attempts to rationalize the extra parameters in terms of the energetics of melting [23] (providing an enhanced value for $T_m \Delta S_m$).

Empirical approaches

Extending the empirical approach can produce methods with good accuracy and excellent computational performance. Empirical methods treat the problem of predicting solubility as a pure QSPR. The response variable (Y) is the measured solubility and the compounds are described using a combination of molecular descriptors (X). Some form of regression is applied to relate Y to X and the result is the solubility model. Within a tight chemical series, Free-Wilson regression [24] can be used to squeeze more mileage from limited experimental results and this is something that Syngenta physical chemists do quite often with good results. Methods that are more general require descriptors and regression techniques that are more sophisticated, the combination of which gives each technique its flavour. The challenge is to model a response that is distinctly non-linear, which implies either that a non-linear regression method is employed or that the molecular descriptor is sufficiently flexible and complex to cope. Both paths create statistical minefields that must be negotiated with some care. Selecting a subset of descriptors from a larger pool (often done iteratively with model building) can render standard statistical measures (such as the F-value) meaningless as the number of degrees of freedom is effectively hidden [25]. The usual way to guard against over-optimistic statistics derived from the training set is to use an independent test set. Once a model has been derived, it is tested against data not used to train it and these results should give a better indication of how useful the model will be for predicting new compounds. The size of the test set varies quite widely (typically 10–50% of the total available data for this sort of study) since holding back data for blind testing reduces the amount of data available for training the model. One way out of this bind is to use leave-one-out (LOO) cross-validation [26] where each compound in turn is set aside, a model built with the rest of the compounds and used to predict a value for the compound omitted. The pool of predictions forms a pseudo-independent test set which is the same size as the training set. This makes efficient use

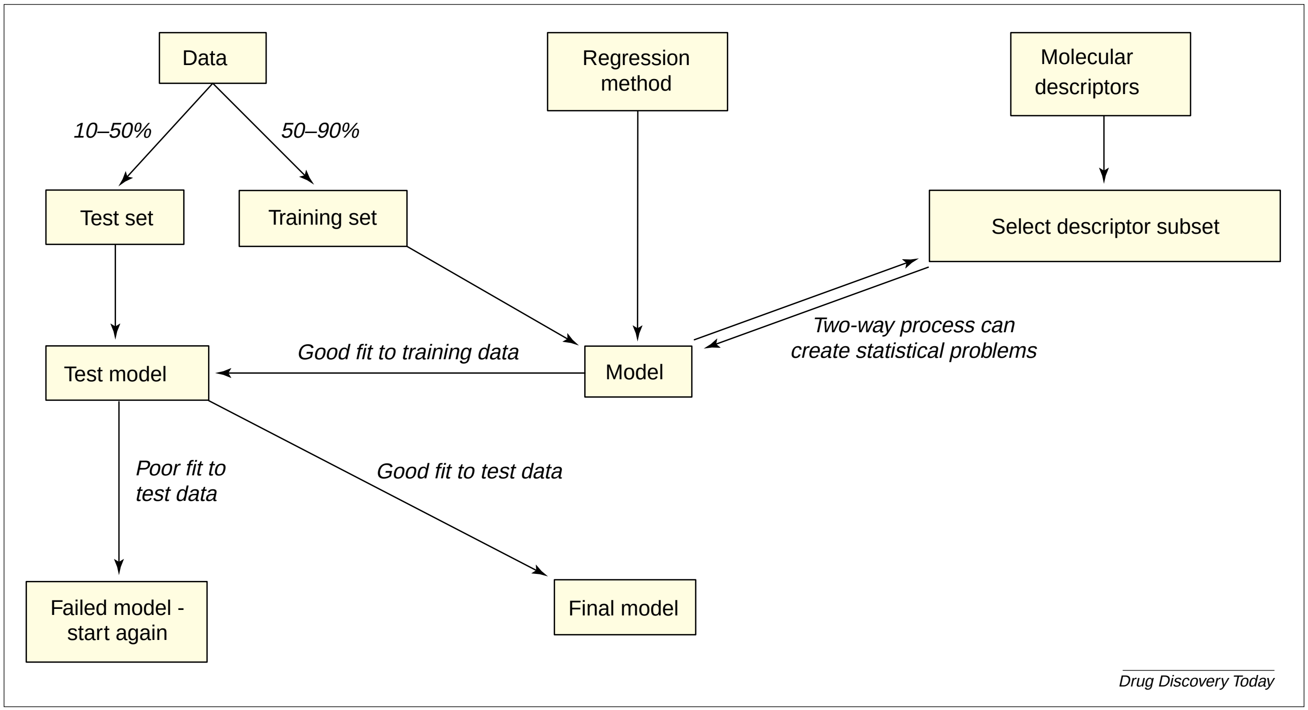

FIGURE 2

The workflow involved in generating a QSPR model for aqueous solubility. The investigator starts with solubility data, a regression method and a description of the molecules. The data is split into a training set which is used to produce the model and a test set which is set aside for validating the final model. The training data, regression method and descriptor subset are combined to produce a model with a good fit to the training set. Part of the process of building this model usually involves some sort of descriptor selection (hence the two-way arrow), a procedure which can make the initial model statistics over-optimistic. The model is then validated against the test set – a good fit to the test data indicates the process has been successful, a poor fit may mean the investigator needs to reconsider their choice of regression method or molecular descriptor.

of the data while providing some sort of blinded evaluation of the model. A schematic of the model building process is shown in Figure 2.

The first choice facing the investigator is how to describe the molecule. There are a large number of 2D, 3D and whole molecule descriptors that can be calculated for a compound, so which ones do we pick? The choice can be guided by the sorts of parameters that are known to be important for interactions between water and solute as discussed earlier or we can use descriptors that allow the similar property principle (similar molecules have similar properties given the right measure of similarity [27]) to be applied. The log P based methods and Abraham's Linear Solvation Energy Relationship (LSER) method [28] tend towards the former approach with a relatively small number of parameters accounting for hydrogen bonding (to solvent and within crystal) and solvent cavity formation. Butina and Gola used an artificial intelligence (AI) technique [21] to produce a small number of rules and local models based on a proprietary log P estimate and additional structural terms. Cheng and Merz used a genetic algorithm to select variables for a linear model [22], with Alog P98 [29] (a log P estimation method with broadly similar accuracy to CLOGP) being the most significant contributor. The author has produced a method called ESOL, which has only four parameters yet reasonable predictive performance on drug-sized molecules [23].

The similar property principle has been used successfully in several approaches, chiefly group [30–33] or atom [34,35] contribution methods and techniques based on 2D connectivity [36–38]. Group contribution methods work in a similar way to well-established programs like CLOGP and ACD/logD, which calculate log P as a sum of fragment contributions from the molecule. The compound is broken down into a series of substructures, each of which has a value associated with it – the solubility is simply the sum of these values. The method works rather well for log P, but this process merely involves the movement of compound between two liquid phases. Solubility crucially involves a change from solid to liquid, and this makes it harder to separate the contributions of individual parts of a molecule to the whole process. Interactions between fragments can be incorporated into group contribution methods via correction factors but this can quickly get unwieldy. Nevertheless, this type of method can be very quick to run on large numbers of compounds and can produce reasonably accurate results.

The 2D chemical diagram is an information rich representation of a molecule, but one that can be hard to shoehorn into a regression method. Topological descriptors [39] are one way of taming the chemical graph as they are directly derived from the connectivity of a molecule and present the information in a manageable, fixed length form. The work of Huuskonen [40–43] best exemplifies

TABLE 1.

Comparison of methods in terms of complexity, predictive ability and throughput

Lead author	Regression/ modeling method	Descriptors	Number of parameters in model	Number in training set	Number in test set	Standard Error (test set)	Standard error (21 common compounds [32])	Throughput of final model
Huuskonen [36]	ANN[a]	2D topological	30	884	413	0.60	0.55	High
Hou [35]	LR[b]	Atomic	78	1290	120	0.79	0.70	High
Jorgensen [11]	LR and MC simulation[c]	Whole molecule	5	150	149[d]	0.72	0.73[e]	High
Yan [47]	ANN	3D descriptors	40	797	496	0.59	0.77	Medium
Wegner [54]	GA[f] and ANN	2D topological	9	1016	253	0.54	0.84	High
Delaney [23]	LR	Whole molecule	4	2874	528	0.96	0.78	High
Klopman [32]	LR	2D substructural	118	1168	120	0.79	0.83	High
Liu [37]	ANN	2D topological	7	1312	258	0.72	0.87	High
Butina [21]	AI[g]	Whole molecule	52	2688	640	1.01	0.82[h]	High
Klamt [15]	LR	Quantum mechanical	3	150	107	0.61	0.91[e]	Low

[a]ANN, artificial neural network. [b]LR, linear regression. [c]MC simulation, Monte Carlo simulation. [d]Leave-one-out cross-validation rather than blind test set. [e]Only 13 from 21 results reported. [f]Genetic algorithm (GA). [g]Artificial intelligence (AI). [h]Only 11 from 21 results reported.

the group of empirical methods that employ topological descriptors with a supervized neural net [44–46]. The model that emerges from training a neural net may be interrogated by applying new input parameters but the raw model itself (a set of connection weights essentially) is not very useful. Topological parameters can be rather abstruse, further restricting insight into the problem. What can be said is that these types of model achieve exceptionally good empirical performance (down to the experimental accuracy of the training set), and if not overtrained, can be applied generally. Once trained these models execute quickly, making profiling of large libraries or collections feasible.

Given the success of 2D descriptors it might be supposed that 3D descriptors would be even better as they should reflect physical reality more faithfully. There are relatively few examples of 3D empirical methods being successfully deployed in the literature, reflecting the difficulties in finding alignment free descriptors that are sufficiently discriminating. Gasteiger and Yan [47] have produced a 3D method that performs at about the same level as typical group contribution or log P based methods, but the extra effort involved in producing the descriptors begs the question as to why one might use them [48]. Clark and co-workers have developed a QSPR method [49,50] based on an accurate representation of the charge distribution within a molecule (derived from a QM calculation) and a neural network.

Data regression methods

All empirical methods depend on some form of regression to build a predictive model. The best choice of regression method depends largely on the number and nature of the descriptors used. The right pairing is important, as the final model has to capture the non-linear nature of the problem. The types of regression that have been employed include multiple linear [51], partial least squares [52,53], artificial neural nets [36,37,38,40,44,54] (ANNs), support vector machines [55], Bayesian neural networks [56], genetic algorithms [54] and clustered regression [21]. The relative popularity of supervized ANNs reflects a situation where a non-linear response is appropriate and relatively large quantities of training data are available. Not atypically, they show excellent empirical performance at the cost of a certain opacity and a limited ability to generalize – ANNs are the archetypal black-box method. Simple linear regression can also be used if the right descriptors are available, and has the advantage of producing interpretable models with well-understood statistical performance. Problems associated with variable selection and non-linearity can be addressed using more elaborate methods such as partial least squares [52], principle components regression [57] or genetic algorithms [22].

Practical implementation

Regardless of the method used to predict solubility, some issues may impact on its effectiveness. Any predictive method relies on the correct representation of the molecules under scrutiny. This may seem like a trivial point, but it has been our experience that one should not always believe that the structure in a database is a faithful reflection of the true state of the sample in the tube! Apart from gross mistakes in registering the structure, there are issues around the tautomeric form of the molecule. Some companies enforce rules for representing tautomers in corporate databases, which ensures consistency for indexing purposes but can be misleading if used as a physical representation in a prediction program. We feel that correcting the representation before prediction is a necessary step. The compound training sets used to produce most of the models in the literature are subject to an unknown degree of experimental error (up to half a log unit [28]) and also tend to over-represent compounds with low molecular weights [58,59]. The error performance of predictive methods is also interesting, as the largest errors seem to occur with low solubility compounds – that is, the ones we are often most interested in getting accurate values for. This unfortunate behaviour is not well reflected in the overall statistics usually quoted to support the accuracy of the model. An element of 'fit-for-purpose' has to be applied when assessing these methods. If the aim is to get accurate predictions for a small series of difficult compounds, the investigator should probably not rely on one method and should attempt to incorporate any measured results. On the other hand, if all that is required is a solubility distribution or profile for a large number of compounds then a single, fast method, free from systematic bias is probably appropriate. Inaccuracy in the exact solubility value at the low solubility end of things may be excusable if all you are trying to do is remove 'brickdust' from a collection. Table 1 attempts to summarize some of the reported methods in terms of their parameters, regression method, predictive throughput, reported test set performance and their performance against a common set of compounds (21-molecule set found in many solubility prediction papers [32]). All of the methods managed to achieve standard errors of less than one for the 21-member, common test set.

Most of the solubility estimates presented in this review only work reliably on non-charged compounds, they are designed to reproduce the intrinsic solubility of a compound. Some methods (notably COSMO-RS and ACD/LogD) allow solubility to be determined at different pHs, salt concentrations and even with different solvents. This last point is of particular interest to high throughput screeners as DMSO is the most commonly used solvent for their assays. The implications of solubility for a compound's 'screenability' in HTS can be profound, especially when amorphous samples that appear to have dissolved crash out days later through Ostwald ripening [60]. Accurate DMSO solubility prediction is in its infancy [61,62], largely owing to a lack of publicly available solubility data, but could be the focus of future work in this field. Pharma Algorithms have developed a DMSO solubility prediction method using private data obtained from Specs (a commercial compound vendor), which shows promise.

Conclusions

The current state-of-the-art in aqueous solubility calculations offers a wide range of methods that can produce results close to experimental accuracy for uncharged, small compounds (most of the methods discussed should produce solubility estimates with mean absolute errors comfortably less than 1 log unit). There remain areas where improvement would be welcome, particularly when dealing with real compounds of practical interest. However, most prediction methods rely on publicly available data for training and validation, so the relevance issue should not be laid at the door of the people developing methodology. Publishing better data might allow better methods to be developed [59,63]. Many of the methods can be deployed widely to bench chemists as they simply use the structure diagram of the compound as input (e.g. Syngenta chemists can use ESOL via a web-interface), which should make late-stage solubility problems less of a feature of the development landscape. There remain some interesting challenges within the area concerning solubilities at different pHs (charged compounds), and different (and mixed) solvent systems such as DMSO and DMSO/water.

Perhaps the most interesting aspect of this field of study is the sheer diversity of approaches that have been brought to bear. Aqueous solvation in some ways seems like a relatively straightforward process to model (particularly when compared to biological systems), yet no one prediction method could be called definitive. The ideal technique would be accurate, but also robust and comprehensive, coping with small, model compounds and large, drug-like molecules alike. It would execute quickly on large libraries of virtual compounds, yet be capable of providing insight into the causes of poor solubility in individual cases. As long as this ideal remains elusive, researchers will continue to be drawn to the problem.

References

1 Musha, D. (1956) Body water in man. I. Total body water in normal subjects and edematous patients. *Tohoku J. Exp. Med.* 63, 309–317

2 Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23, 3–25

3 Clarke, E.D. and Delaney, J.S. (2003) Physical and molecular properties of agrochemicals: An analysis of screen inputs, hits, leads and products. *Chimia (Aarau)* 57, 731–734

4 Kariv, I. *et al.* (2002) Improvement of "hit-to-lead" optimization by integration of *in vitro* HTS experimental models for early determination of pharmacokinetic properties. *Comb. Chem. High Throughput Screen.* 5, 459–472

5 Yang, G. *et al.* (2002) Prediction of the aqueous solubility: comparison of the general solubility equation and the method using an amended solvation energy relationship. *J. Pharm. Sci.* 91, 517–533

6 Valko, K. (2004) Application of high-perfomance liquid chromatography based measurements of lipophilicity to model biological distribution. *J. Chromatogr. A.* 1037, 299–310

7 Morris, J.J. and Bruneau, P.P. (2000) Prediction of physicochemical properties. In *Virtual Screening for Bioactive Molecules* (Bohm H.G. and Schneider G., eds), pp. 33–58, Wiley-VCH

8 Popa-Burke, I.G. et al. (2004) Streamlined system for purifying and quantifying a diverse library of compounds and the effect of compound concentration measurements on the accurate interpretation of biological assay results. *Anal. Chem.* 76, 7278–7287

9 Clark, D.E. and Grootenhuis, P.D.J. (2002) Progress in computational methods for the prediction of ADMET properties. *Curr. Opin. Drug Discov. Devel.* 5, 382–390

10 Sanghvi, T. *et al.* (2003) Estimation of aqueous solubility by the general solubility equation (GSE) the easy way. *QSAR & Combinatorial Science* 22, 258–262

11 Jorgensen, W.L. and Duffy, E.M. (2000) Prediction of drug solubility from Monte Carlo simulations. *Bioorg. Med. Chem. Lett.* 10, 1155–1158

12 Kier, L.B. *et al.* (2001) Cellular automata models of aqueous solution systems. In *Reviews in Computational Chemistry* (Vol. 17) (Lipkowitz, K.B. and Boyd, D.B., eds), pp. 205–254, Wiley-VCH

13 Gardner, M. (1970) The fantastic combinations of John Conway's new solitaire game "life". *Sci. Am.* 223, 120–123

14 Cramer, C.J. and Truhlar, D.G. (1995) Continuum solvation models: Classical and quantum mechanical implementations. In *Reviews in Computational Chemistry* (Vol. 6) (Lipkowitz, K.B. and Boyd, D.B., eds), pp. 1–72, Wiley-VCH

15 Klamt, A. (2002) Prediction of aqueous solubility of drugs and pesticides with COSMO-

RS. *J. Comput. Chem.* 23, 275–281

16 Chickos, J.S. *et al.* (1991) Estimating entropies and enthalpies of fusion of organic compounds. *J. Org. Chem.* 56, 927–938

17 Dannenfelser, R.M. and Yalkowsky, S.H. (1979) Estimation of entropies of fusion of organic compounds. *Industrial and Engineering Chemistry Fundamentals* 18, 108–111

18 Dearden, J.C. (1999) The prediction of melting points. In *Advances in Quantitative Structure–Property Relationships* (Vol. 2) (Charton, M. and Charton, B.I., eds), pp. 127–175, JAI Press

19 Tesconi, M. and Yalkowsky, S.H. (2000) Melting point. In *Handbook of Property Estimation Methods for Chemicals: Environmental and Health Sciences*. (Boethling, R.S. and Mackay, D., eds), pp. 1–27, CRC Press

20 Bergström, C.A.S. *et al.* (2003) Molecular descriptors influencing melting point and their role in classification of solid drugs. *J. Chem. Inf. Comput. Sci.* 43, 1177–1185

21 Butina, D. and Gola, J.M.R. (2003) Modeling aqueous solubility. *J. Chem. Inf. Comput. Sci.* 43, 837–841

22 Cheng, A. and Merz, K.M. (2003) Prediction of aqueous solubility of a diverse set of compounds using quantitative structure–property relationships. *J. Med. Chem.* 46, 3572–3580

23 Delaney, J.S. (2004) ESOL: Estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.* 44, 1000–1005

24 Kubinyi, H. (1993) *QSAR: Hansch Analysis and Related Approaches* (*Methods and Principles in Medicinal Chemistry, Vol. 1*) (Mannhold, R. *et al.*, eds), Wiley-VCH

25 Topliss, J.G. and Edwards, R.P. (1979) Chance factors in studies of quantitative structure–activity relationships. *J. Med. Chem.* 22, 1238–1244

26 Diaconis, P. and Efron, B. (1983) Computer-intensive methods in statistics. *Sci. Am.* 248, 116–130

27 Rouvray, D.H. (1990) The evolution of the concept of molecular similarity. In *Concepts and Applications of Molecular Similarity* (Johnson, M.A. and Maggiora, G.M., eds), pp. 15–42), Wiley

28 Abraham, M.H. and Le, J. (1999) The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. *J. Pharm. Sci.* 88, 868–880

29 Ghose, A.K. *et al.* (1998) Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of ALOG P and CLOG P methods. *J. Phys. Chem. A* 102, 3762–3772

30 Myrdal, P.B. *et al.* (1995) AQUAFAC: Aqueous functional group activity coefficients: Application to the estimation of aqueous solubility. *Chemosphere* 24, 1619–1637

31 Kuhne, R. *et al.* (1995) Group contribution methods to estimate water solubility of organic chemicals. *Chemosphere* 30, 2061–2077

32 Klopman, G. (1992) Estimation of aqueous solubility of organic compounds by the group contribution approach. application to the study of biodegradation. *J. Chem. Inf. Comput. Sci.* 32, 474–482

33 Marrero, J. and Gani, R. (2002) Group-contribution-based estimation of octanol/water partition coefficient and aqueous solubility. *Ind. Eng. Chem. Res.* 41, 6623–6633

34 Sun, H. (2004) A universal molecular descriptor system for prediction of Log P, LogS, LogBB, and absorption. *J. Chem. Inf. Comput. Sci.* 44, 748–757

35 Hou, T.J. *et al.* (2004) ADME evaluation in drug discovery: 4. Prediction of aqueous solubility based on atom contribution approach. *J. Chem. Inf. Comput. Sci.* 44, 266–275

36 Huuskonen, J. (2000) Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.* 40, 773–777

37 Liu, R. and So, S. (2001) Development of quantitative structure–property relationship models for early ADME evaluation in drug discovery: 1. Aqueous solubility. *J. Chem. Inf. Comput. Sci.* 41, 1633–1639

38 Tetko, I.V. *et al.* (2001) Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci.* 41, 1488–1493

39 Hall, L.H. and Kier, L.B. (1991) The molecular connectivity Chi indexes and Kappa shape indexes in structure–property modeling. In *Reviews in Computational Chemistry* (Vol. 2) (Lipkowitz, K.B. and Boyd, D.B., eds), pp. 367–422, Wiley-VCH

40 Livingstone, D.J. *et al.* (2001) Simultaneous prediction of aqueous solubility and octanol/water partition coefficient based on descriptors derived from molecular structure. *J. Comput. Aided Mol. Des.* 15, 741–752

41 Huuskonen, J. (2001) Estimation of water solubility from atom-type electrotopological state indices. *Environ. Toxicol. Chem.* 20, 491–497

42 Huuskonen, J. (2000) Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotopological state indices. *Eur. J. Med. Chem.* 35, 1081–1088

43 Huuskonen, J. and Tetko, I.V. (2000) Application of neural networks for estimating partition coefficient based on atom-type electrotopological state indices. Molecular modeling and prediction of bioactivity, *Proceedings of the European Symposium on Quantitative Structure–Activity Relationships: Molecular Modeling and Prediction of Bioactivity*, Copenhagen, Denmark, 23–28 Aug. 1998, pp. 470–471

44 Yan, A. and Gasteiger, J. (2003) Prediction of aqueous solubility of organic compounds by topological descriptors. *QSAR & Combinatorial Science* 22, 821–829

45 Engkvist, O. and Wrede, P. (2002) High-throughput, in silico prediction of aqueous solubility based on one- and two-dimensional descriptors. *J. Chem. Inf. Comput. Sci.* 42, 1247–1249

46 Yaffe, D. (2001) A fuzzy ARTMAP based on quantitative structure–property relationships (QSPRs) for predicting aqueous solubility of organic compounds. *J. Chem. Inf. Comput. Sci.* 41, 1177–1207

47 Yan, A. and Gasteiger, J. (2003) Prediction of aqueous solubility of organic compounds based on a 3D structure representation. *J. Chem. Inf. Comput. Sci.* 43, 429–434

48 Yan, A. *et al.* (2004) Linear and nonlinear functions on modeling of aqueous solubility of organic compounds by two structure representation methods. *J. Comput. Aided Mol. Des.* 18, 75–87

49 Beck, B. *et al.* (2000) QM/NN QSPR models with error estimation: Vapor pressure and log P. *J. Chem. Inf. Comput. Sci.* 40, 1046–1051

50 Clark, T. (2000) Quantum cheminformatics: An oxymoron? In *Chemical Data Analysis in the Large: The Challenge of the Automation Age* (Hicks, M.G. ed.), Beilstein Institut

51 Zhong, C. and Hu, Q. (2003) Estimation of the aqueous solubility of organic compounds using molecular connectivity indices. *J. Pharm. Sci.* 92, 2284–2294

52 Bergström, C.A.S. *et al.* (2004) Global and local computational models for aqueous solubility prediction of drug-like molecules. *J. Chem. Inf. Comput. Sci.* 44, 1477–1488

53 Wanchana, S. *et al.* (2002) Quantitative structure/property relationship analysis on aqueous solubility using genetic algorithm-combined partial least squares method. *Pharmazie* 57, 127–129

54 Wegner, J.K. and Zell, A. (2003) Prediction of aqueous solubility and partition coefficient optimized by a genetic algorithm based descriptor selection method. *J. Chem. Inf. Comput. Sci.* 43, 1077–1084

55 Lind, P. and Maltseva, T. (2003) Support vector machines for the estimation of aqueous Solubility. *J. Chem. Inf. Comput. Sci.* 43, 1855–1859

56 Bruneau, P. (2001) Search for predictive generic model of aqueous solubility using Bayesian neural nets. *J. Chem. Inf. Comput. Sci.* 41, 1605–1616

57 Gao, H. *et al.* (2002) Estimation of aqueous solubility of organic compounds with QSPR approach. *Pharm. Res.* 19, 497–503

58 Lobell, M. and Sivarajah, V. (2003) In silico prediction of aqueous solubility, human plasma protein binding and volume of distribution of compounds from calculated pKa and Alog P98 values. *Mol. Divers.* 7, 69–87

59 Jorgensen, W.L. and Duffy, E.M. (2002) Prediction of drug solubility from structure. *Adv. Drug Deliv. Rev.* 54, 355–366

60 Lipinski, C.A. (2003) Aqueous solubility in discovery, chemistry, and assay changes. *Methods and Principles in Medicinal Chemistry* 18, 215–231

61 Lu, J. and Bakken, G.A. (2004) Building classification models for DMSO solubility: Comparison of five methods. *Abstracts of Papers, 228th ACS National Meeting*, Philadelphia, PA, USA, August 22–26

62 Balakin, K.V. *et al.* (2004) In silico estimation of DMSO solubility of organic compounds for bioscreening. *J. Biomol. Screen.* 9, 22–31

63 Bergström, C.A.S. *et al.* (2002) Experimental and computational screening models for prediction of aqueous drug solubility. *Pharm. Res.* 19, 182–188