Informal Qualitative Models: A Systematic Approach to their Generation

Adrian Gordon
Laboratoire de Recherche
en Informatique
Batiment 490
Université de Paris-Sud
91405 Orsay, FRANCE

Derek Sleeman **Pete Edwards**
Department of Computing Science
King's College
University of Aberdeen
Aberdeen, AB9 2UE, UK
{sleeman, pedwards}@csd.abdn.ac.uk

Abstract

This paper discusses the concept of an IQM (Informal Qualitative Model) which is seen as a bridge between rigorous, and often intractable, theories on the one hand, and experimental data on the other. We argue that the selection of variables to be explored using quantitative law discovery should be made using background knowledge. However, domain theories are often intractable, and to make progress it is therefore necessary to add assumptions; i.e. one is forced to take particular and often simplifying, perspectives on the domain. IQMs essentially capture these ideas. This paper demonstrates how a set of IQMs for a domain (colligative properties of solutions) can be generated from a base IQM and a set of operators.

Introduction

Early work in the physical sciences involved the investigation of quantitative relationships between variables (such as Newton's and Kepler's laws), and also qualitative relationships, such as objects of type A react chemically with objects of type B to produce objects of types C and D. In many domains, the scientists had to infer both structural and quantitative models, before a full understanding of a phenomenon could be achieved.

Much of Chemistry is concerned with the following questions:

- What substances exist in nature (elements, atoms, ...), and in what structures are they to be found (molecules, polymers, hydrates, ...) ?
- What are the properties of these substances, and structures ?
- What are the mechanisms for combining/breaking apart such substances, structures and sub-structures ?

Previous work in computational scientific discovery has addressed how each of these types of chemical knowledge can be used to elucidate the others. For example, the STAHL and DALTON systems (Zytkow & Simon, 1986; Langley et al., 1987) use the concept of *chemical reaction* to determine the components of a substance (STAHL), and its molecular composition (DALTON). Related systems are REVOLVER (Rose & Langley, 1986, 1988; Rose, 1989), BR3 (Kocabas, 1991) and GELL-MANN (Fischer & Zytkow, 1990, 1992). These systems all use a different mechanism, the use of collisions to produce sub-atomic particles, to establish quark models of the fundamental particles in physics.

Unlike the previous systems, which used only very general heuristics, REVOLVER uses domain specific knowledge in evaluating the models that it generates. GELL-MANN uses *an additivity principle* (in which properties of an object must be the sum of the contributions of the structures from which it is formed) and a *combination and conservation principle* (in which the same fundamental structures must appear on either side of a reaction) to generate and verify models. BR3 is unique in proposing *new quantum properties,* together with a general *conservation of properties* principle, to account for observed particle reactions.

Discovery in a Space of Structural Models

Each of the systems just described assumes that a single mechanism is operating for combining objects, structures or substructures or breaking them apart. In STAHL and DALTON, the mechanism was chemical reaction, in the other systems the mechanism involved the use of high energy collisions to produce sub-atomic particles. Our own earlier work has focused on using a collection of different mechanisms to hypothesise *models of structure*, and on how these models can be used to explain the properties of a physical system; such as the freezing point of an aqueous salt solution, the solubility of an organic

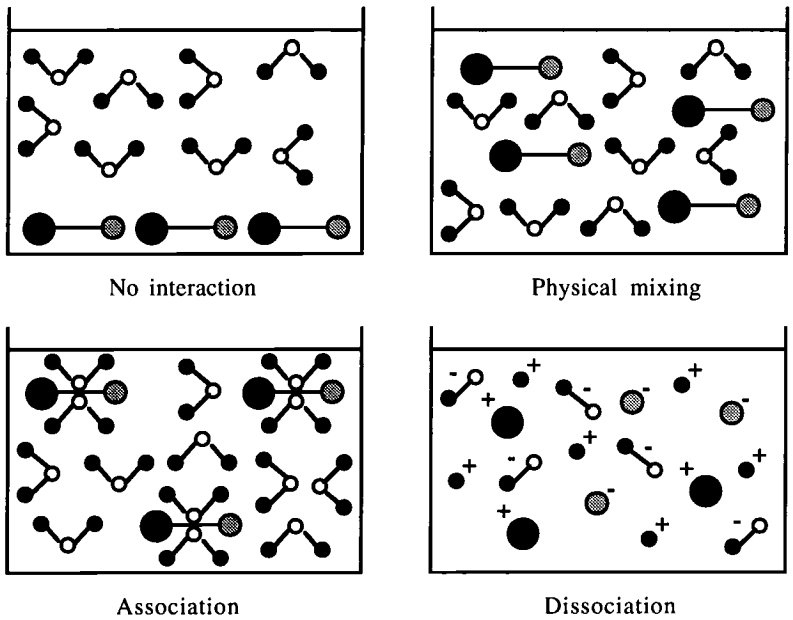

Figure 1 IQMs for Solution Chemistry

compound, or the chemical shift of an atom in a ^{13}C NMR spectrum.

Informal Qualitative Models

The types of structural models we are concerned with have been termed Informal Qualitative Models, or IQMs (Sleeman et al., 1989). They are *informal* and *qualitative* in the sense that they cannot be directly verified by observation. For example, they cannot be verified by the presence or absence of particular reactions. A critique of the BACON family of programs (Langley et al., 1987) led us to articulate the concept of IQMs. Besides being concerned about noise-handling capabilities in BACON, we were exercised by the fact that such quantitative law discovery systems needed to be given the *dependent* and *independent* variables, and their associated sets of values. Sleeman et al. (1989) argued that one of the most challenging tasks for a scientist was to decide which variables to explore, and then to design appropriate experiments to gather the data (which could be passed to a quantitative law discovery program for analysis). We were also aware that in most scientific domains some background theory exists, which is generally used to inform the choice of experimental variables and the likely form of the relationships between them. Additionally, we were also aware that, in many areas of science, the theory is either intractable for all but trivial cases (as in Quantum Mechanics) or too abstract to have any real predictive power (as is seen in the early history of solution chemistry). In order to produce useful instantiations of such theories it is necessary to introduce a series of plausible assumptions. Informal Qualitative Models are one source of such assumptions. Given a series of IQMs, we argued that it would be possible to use these to interpret existing data, or alternatively to design experiments to distinguish between the IQMs. Once the experimental data is available, it is analysed from the perspective of each of the IQMs which specify the relationships/instantiations for variables in the model; then a law discovery mechanism can be used to infer numerical relationships. This process is repeated for each of the IQMs and the one which produces the set of (quantitative) equations which is deemed to be the most plausible/ acceptable is said to be the appropriate IQM (Gordon et al., 1994). For example, a careful analysis of the history of early solution chemistry research shows that progress in the understanding of this domain came about by the proposal of a set of increasingly elaborate models. Interpreting the same experiments from the perspective of each of these models led to the formulation of a set of different quantitative laws. Applying a different model could result in a quantitative law which was preferred because it was more nearly linear, or resulted in fewer exceptions or anomalies (Gordon, 1992, 1993, *in preparation*).

Our earlier work on IQMs (Sleeman et al. 1989; Stacey, 1992; Gordon, 1992, 1993) suffered largely from the *ad hoc* nature of the models, which were simply considered to exist in a scientific domain *a priori*, and be available for use by a scientist, with little consideration of their origin,

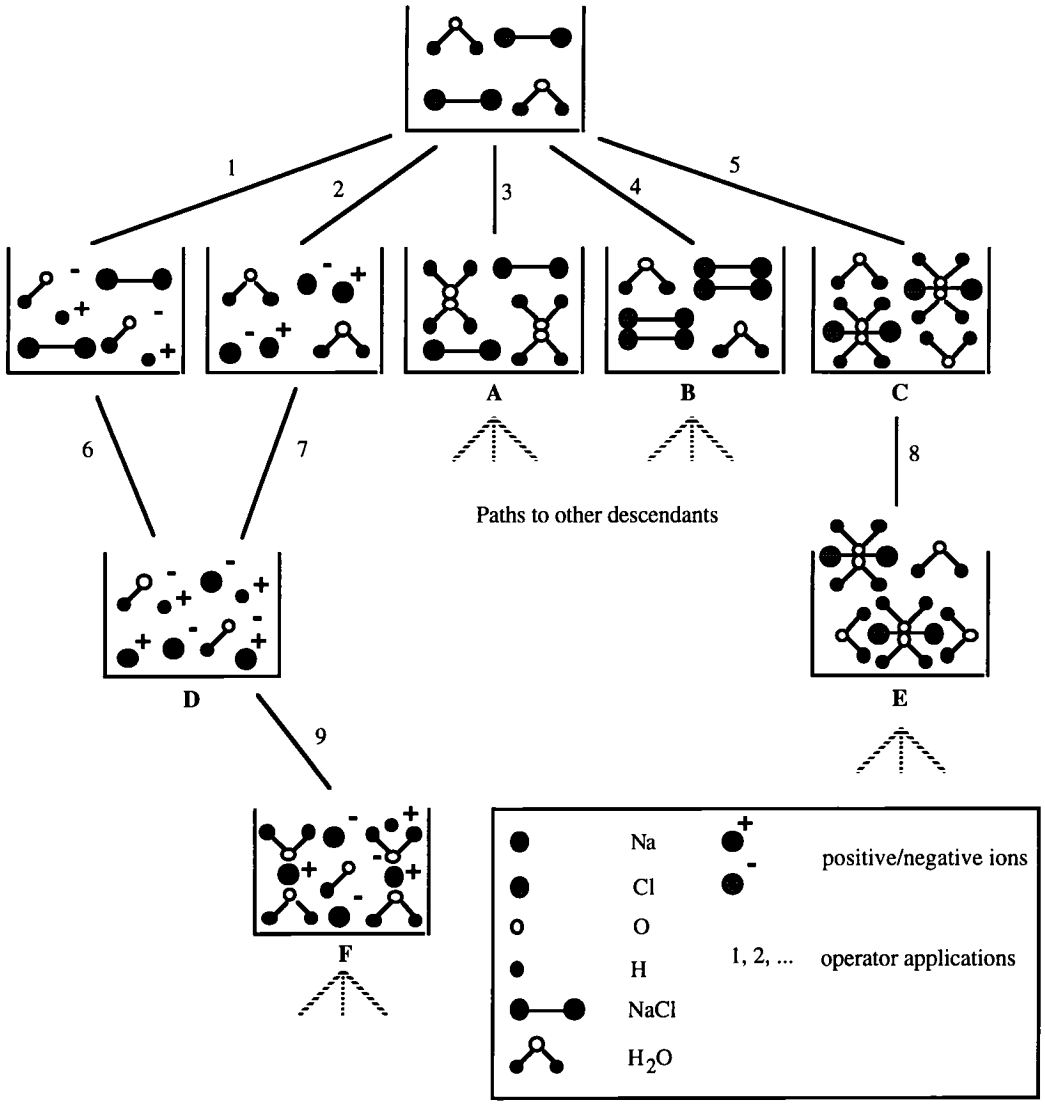

Figure 2. A Partial search space of solution chemistry models

or the obvious relationships amongst them. Figure 1, for example, which is adapted from Sleeman et al. (1989), shows four examples of IQMs which can apply to the domain of solution chemistry.

In Figure 1, the *No Interaction* model represents the case where particles of solute and solvent do not interact at all in a solution. *Physical Mixing* represents solutions in which particles of solute and solvent are uniformly physically distributed throughout the solution, but neither of them are changed chemically in any way. The third model represents an *Association* between solute and solvent particles. Particles of solute are associated with particles of the solvent in a fixed ratio of numbers of molecules. An example would be a solution consisting of *hydrated* salt particles dissolved in water. The figure illustrates the case in which two molecules of water are associated with each molecule of salt. In the final model, *Dissociation*, both solute particles and some of the solvent particles are dissociated into their constituent ions (represented by the charge signs on the particles in the figure). Similar sets of models have been elaborated in other domains, such as ^{13}C Nuclear Magnetic Resonance, the solubility of organic compounds, and celestial mechanics.

An important advance on this work, considers IQMs far more systematically (Gordon et al., 1994). In this approach, the IQMs in a domain are generated from a "root" model, the simplest model possible in a domain, by the application of a set of *model generation operators*. A search space of models can thus be generated (Newell & Simon, 1972). Figure 2 shows a partial search space for

the solution chemistry domain, specifically for aqueous solutions of common salt.

In Figure 2, the root model for the IQM search space is the *Physical Mixing Model* already seen in Figure 1, in which the solute and solvent are simply physically mixed with one another in the resulting solution, with neither of them chemically changed in any way. Subsequent models are generated from this original model by the application of a set of operators. One such operator, **combine-nonionic**, takes five arguments:

combine-nonionic(*model, object1, object2, ratio, degree*)

These arguments are an existing ***model***, two ***objects*** which are to be combined in the resulting model, a ***ratio***, which represents the number of instances of ***object2*** which are to be associated with each instance of ***object1*** in the resulting solution, and a ***degree*** (between 0 and 1) indicating the extent of application of the operator. For example, operator application 5 in Figure 2, would be instantiated as follows to generate model *C* :

combine-nonionic(root, NaCl, H_2O, 2, 1)

That is, the operator is applied to the model **root** (the simplest in the domain); it combines a molecule of common salt with two molecules of water; and it applies to all salt molecules in solution. Essentially, we are saying that in the model of solutions generated by this operator, all salt molecules exist in solution associated with two molecules of water. This *Single Association* model of solutions is due historically to Rüdorff.

The other labelled models in Figure 2 are as follows: Model *A* indicates that all molecules of solvent are associated with another solvent molecule. This model was first proposed historically by Raoult. Model *B* proposes that molecules of the solute are associated with one other solute molecule in solution, this model is also due to Raoult. Model *E*, the *Multiple Associations* model, due historically to De Coppet, proposes that salts can exist in solution in two different states of hydration simultaneously. A second hydrate is shown, with salt and water molecules associated in the ratio 1:4. Model *D*, the basis of the current model of solutions, is the *Ionic Dissociation* model. This model proposes that, in solution, solute and solvent are dissociated into their constituent ions, and is due to Arrhenius. Model *F* is in a sense a hybrid of two of the models we have seen so far; it was proposed to account for some of the anomalies which remained unexplained even by Arrhenius' model, and suggests that ions in solution can themselves be associated with water molecules.

Further details of the history of solution chemistry, and the operators used to generate each of the models in Figure 2 are to be found in Gordon et al. (1994) and Gordon (*in preparation*). Additionally, Gordon (1993) demonstrates how HUME, a model-driven discovery system, applies the single association model to the problem of law discovery in solution chemistry. The goal in this case is to discover numerical laws that can describe the behaviour of the freezing points of aqueous salt solutions. The application of the *Single Association* model to this problem can lead to the discovery of more satisfactory laws than is possible by applying the simpler *Physical Mixing* model.

Although we have now established a more systematic approach to Informal Qualitative Models, by showing how a search space such as that of Figure 2 can be generated from a single, root model for the domain, several questions remain. Most importantly, of course, is the question of the source of this most primitive model, and of the model-generating operators applied to it. These are the challenges for the next phase of our work.

References

Fischer, P.J. and Zytkow, J.M. (1990). Discovering Quarks and Hidden Structure. *Methodologies for Intelligent Systems 5*, New York: North Holland, 363-370.

Fischer, P.J. and Zytkow, J.M. (1992). Incremental Generation and Exploration of Hidden Structure. In *Proceedings of the ML92 Workshop on Machine Discovery*, J. M. Zytkow (Ed.), 103-110.

Gordon, A. (1992). Informal Qualitative Models and Scientific Discovery. In *Proceedings of the ML92 Workshop on Machine Discovery*, J. M. Zytkow (Ed.), 98-102.

Gordon, A. (1993). Informal Qualitative Models and the Depression of the Freezing Point of Solutions. In *Working Notes for the MLnet Workshop on Machine Discovery*, P. Edwards (Ed.), 56-60.

Gordon, A. (in preparation). *Informal Qualitative Models in Scientific Discovery*. Thèse de Docteur en Sciences, Université de Paris-Sud, Centre D'Orsay.

Gordon, A., Edwards, P., Sleeman, D. and Kodratoff, Y. (1994). Scientific Discovery in a Space of Structural Models: An Example from the History of Solution Chemistry. In *Proceedings of the 16th Conference of the Cognitive Science Society*, 381-386.

Kocabas, S. (1991). Conflict Resolution as Discovery in Particle Physics. *Machine Learning*, 6, 277-309.

Langley, P., Simon, H.A., Bradshaw, G.L. and Zytkow, J.M. (1987) *Scientific Discovery: Computational Explorations of the Creative Processes*, Cambridge, MA: MIT Press.

Newell, A. and Simon, H. A. (1972). *Human Problem Solving*, Englewood Cliffs, NJ: Prentice-Hall.

Rose, D. and Langley, P. (1986). Chemical Discovery as Belief Revision. *Machine Learning*, 1, 423-452.

Rose, D. and Langley, P. (1988). A Hill-climbing Approach to Machine Discovery. In *Proceedings of the Fifth International Conference on Machine Learning*, 367-373.

Rose, D. (1989). *Belief Revision and Scientific Discovery*. Doctoral Dissertation, Department of Information and Computer Science, University of California, Irvine.

Sleeman, D.H., Stacey, M.K., Edwards, P. and Gray, N.A.B. (1989). An Architecture for Theory-Driven Scientific Discovery. In *Proceedings of the Fourth European Working Session on Learning*, 11-24.

Stacey, M.K. (1992). *A Model-Driven Approach to Scientific Law Discovery.* Ph.D. Thesis, Department of Computing Science, University of Aberdeen.

Zytkow, J. M. and Simon, H. A. (1986). A Theory of Historical Discovery: The Construction of Componential Models. *Machine Learning*, 1, 107-137.