Data Mining via ILP: The Application of Progol to a Database of Enantioseparations

Christopher H.Bryant[1]

School of Computing and Mathematics, University of Huddersfield, HD1 3DH, UK. **

Abstract. As far as this author is aware, this is the first paper to describe the application of Progol to enantioseparations. A scheme is proposed for data mining a relational database of published enantioseparations using Progol. The application of the scheme is described and a preliminary assessment of the usefulness of the resulting generalisations is made using their accuracy, size, ease of interpretation and chemical justification.

1 Introduction

This paper describes a scheme for performing data mining on a chemical database and makes a preliminary assessment of the results of applying the scheme. The scheme utilises Progol, a domain independent ILP tool which is available in the public domain. As far as this author is aware, this is the first paper to describe the application of Progol to enantioseparations.

An enantioseparation [11] is the separation of two enantiomers. In order to perform an enantioseparation a chiral selector must be used which has a preference for one of the enantiomers in the pair as a consequence of its stereochemistry. This is usually achieved by selecting a suitable Chiral Stationary Phase (CSP).

The main areas to which ILP has been applied previously [1] are scientific discovery, knowledge acquisition and programming assistants. Applications of ILP to scientific discovery and knowledge acquisition include drug design, protein folding, diterpene structure elucidation from ^{13}C NMR spectra [7], diagnosis of faults in the power supply of satellites and rheumatology diagnosis. Work conducted as part of the project described in this paper applied Golem to enantioseparations [4].

2 Drug Separation Data

The research described in this paper used data that was taken from a recent study [6] that investigated the ability of seven CSP chiral selectors to separate enantiomeric drugs. The training data set [2] contains data on 197 separations involving 50 drugs whose structures vary widely.

** c.h.bryant@hud.ac.uk Tel: +44-1484-473047 Fax: +44-1484-421106
URL: http://www.hud.ac.uk/schools/comp+maths/private/chb/top.html

The data was downloaded from a relational database of enantioseparations which uses attributes to represent chemical structural features of enantiomers. (For a description of the design of the database see [2].) Names of attributes in the database represent occurrences of chemical features and the values of these attributes represent the distances of the chemical features from the chiral centre in terms of the number of connecting bonds. There are three attributes for each chemical feature represented in the database. These represent the first, second and third occurrence of a feature. (Full details of the representation of chemicals in the database are given in both [3] and [2].)

3 The Induction Scheme

This section describes a scheme for the induction of the generalisations needed for recommending a suitable CSP chiral selector for a given enantiomer pair.

3.1 Why Progol was Selected

A non-interactive, non-incremental ILP tool was sought. The use of an interactive ILP tool was precluded because there was no suitable expert available to act as an oracle. The use of an incremental tool was unnecessary because all of the data was available prior to induction.

The three most widely field tested tools of this type are FOIL [10], Golem [9] and Progol [8]. Golem and Progol were preferred to FOIL because they have been applied successfully to chemical domains previously. Progol[3] was used rather than Golem because it does not suffer from some of the limitations of Golem such as the prohibition of non-ground unit clauses in the input files and the restriction to including only determinate clauses.

3.2 Knowledge Representation

One of the aims of the scheme is to induce generalisations that will suggest which CSP chiral selector should be used to separate a given enantiomer pair. To represent such generalisations using first order logic it is necessary to use a predicate that maps enantiomer pairs to CSP chiral selectors. Hence the predicate separates_on(E, C), where E=enantiomer pair and C=CSP chiral selector, is used to represent the separations in the data. The separates_on literals are divided into two groups, positives and negatives, which reflect whether the separation they represent is successful.

The bias of Progol gives rise to a choice of four options for representing the relationships between the data on enantioseparations and the data on chemical features of enantiomers. In each option D = distance from the chiral centre.

1. **has_feature(F, E, D)** F = chemical feature (including the occurrence).

[3] Version C4.1 of Progol was used in this project.

2. **has_feature(F, O, E, D)** Here F = chemical feature and O = occurrence of a chemical feature.
3. **<chemical_feature>(E, D)**
4. **<chemical_feature>(O, E, D)**

The second and fourth options require the names of the chemical feature attributes to be split into their constituent feature and occurrence parts. The third and fourth options require a predicate for each of the chemical feature attributes.

The induction scheme described in this paper uses the second choice for reasons explained in Sect. 3.3. It is interesting to compare Progol with Golem in this respect. Golem only allows the third predicate because Golem is restricted to inducing determinate literals.

(Obviously the language bias of Progol would allow many other predicates that represent the chemical features of enantiomers, not least because it accepts non-ground unit clauses.[4] However, since the approach taken in this project is to develop rules from data stored in the database, the only predicates considered are those for which instantiations can be generated by downloading and reformatting data from the database.)

3.3 Generalising Distances, Occurrences and Features

Enabling Progol to Generalise Distances. Providing Progol with just those predicates selected in Sect. 3.2 is not sufficient to enable it to make useful generalisations about the domain because, without additional background predicates, Progol is not able to make generalisations about the distance at which the chemical features of an enantiomer pair must occur in order that the pair be separated by a given CSP chiral selector. Without additional background predicates Progol will only induce clauses that reason about the presence of chemical features at particular distances or at no particular distance. This author believes that for a machine induction tool to be of use for enantioseparations it must be able to generalise distance values in a more flexible manner than this. This section describes the component of the scheme that allows Progol to make such generalisations. The generalisations are expressed as clauses of the form shown below where `gd` is a constant representing a merge of distance values.

separates_on(E, C):- has_feature(F, O, E, gd).

An example of this form of clause is:–

separates_on(E, dnbpg):– has_feature(bg6, third_closest, E, one_or_two).

where bg6 represents a six-membered aromatic ring and dnbpg represents the chiral selector (R)-N-(3,5-dinitrobenzoyl)-phenylglycine.

[4] The separate issue of using non-ground unit clauses to represent background chemical knowledge is discussed later in this paper.

If the background knowledge includes a series of clauses of the form shown below, where the constant `gd` is a generalisation of another constant `sd`, and the modes[5] shown below[6] are declared then Progol is able to generalise distances.

```
has_feature(F, O, E, gd):- has_feature(F, O, E, sd).
:–modeh(1,separates_on(+enantiomer_pair,#csp))?
:–modeb(500,has_feature(#feature,#occurrence,+enantiomer_pair,#distance))?
```

Enabling Progol to Generalise Occurrences. The previous section described the component of the scheme that enables Progol to make useful generalisations about the distance of chemical features from the chiral centre. In the absence of any clauses in the background knowledge for generalising occurrences and given the modes declared, Progol was restricted to inducing clauses that reason about the presence of *particular occurrences of* chemical features. This author believes that in some cases it may not matter whether a chemical feature is the closest, second closest or third closest occurrence of that feature, as long as the feature is present at a particular distance or within a range of distance values. If a machine induction tool is to be able to induce clauses from the database that reflect this then it must be capable of generalising the data on both the occurrences and the distances. Progol can be given this capability by declaring that the term occurrence can be either – or #. Of course, this assumes that there is a term representing the occurrence data. Thus this component of the scheme requires that either the second or fourth choice predicate for representing features (described in Sect. 3.2) is used to represent the relationships between the data on enantioseparations and the data on chemical features of enantiomers. When the predicate has_feature(F, O,E, D) is used the mode declarations listed in Sect. 3.3 are supplemented by the one shown below. This enables Progol to reason about particular occurrences of chemical features or any occurrence of chemical features.

```
:–modeb(500,has_feature(#feature,–occurrence,+enantiomer_pair,#distance))?
```

Enabling Progol to Generalise Features. Chemists often reason in terms of chemical features that are more general than those that are represented in the database. For example, reasoning about features such as aromatic rings or carbonyl groups is common place in chemistry but these features are not represented in the database. Enabling a machine induction tool to generalise the data

[5] modeh/2 and modeb/2 describe the 'forms' of literals that are allowed in the head and body respectively of a hypothesised clause. The first term, referred to as the recall number, specifies an upper bound on the number of successful calls to a predicate. The second term declares the mode and type of each term of the predicate. Types may be unary predicates defined in the background knowledge. Modes are either input (+), output (-) or constant (#).

[6] The value required for the recall number was determined empirically by increasing the verbosity of Progol.

in the database on the chemical features would give it the potential to generate more concise clause-sets and to make discoveries that would not be possible otherwise.

Many of the chemical features represented in the database have more general chemical features in common and some of these, in turn, have yet more general features in common. The relationships between these features can be represented in first-order predicate logic using the two clauses shown below and ground instantiations of the isa predicate. (Note that is_a and isa are two different predicates.) It must be emphasized for those readers not familiar with the domain that these ground unit clauses represent concepts that are omnipresent in chemistry.

is_a(A, B):– isa(A, B).   is_a(A, C):– isa(A, B), is_a(B, C).

This requires that the features are represented by a term; thus this component of the scheme requires that the first or second choice of predicate for representing features (see Sect. 3.2) is used. Since the component of the scheme for generalising occurrence data requires that the second or fourth choice of predicate is used, the scheme uses the second choice, namely has_feature(F, O, E, D).

4 Results of Applying the Scheme to the Data Set

Progol induced clauses for five of the seven CSP chiral selectors. The clauses are all short: they have either one or two literals in their body. This makes it easy to interpret them and to understand why particular separations are covered by particular clauses.

Consider the clauses for (R)-N-1-(α-naphthyl)ethylaminocarbonyl-(S)indoline-2-carboxylic acid which are shown in Fig. 2 together with their English translation. Clause *a* covers eleven of the 21 successful separations on the CSP chiral selector mentioned. When the clause is considered in conjunction with the structures of the enantiomers covered by the clause it becomes apparent that the clause represents the fact that the enantiomers fit the structural template shown in Fig. 1.

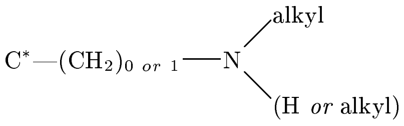

Fig. 1. Structural template of enantiomers from which clause *a* was induced.

Clause *b* covers five more of the successful separations on the selector and clause *c* another two. Again it is clear why the separations are covered by the clauses. The structural features referred to in the clauses are easily discerned on the structure diagrams: both clauses refer to ring features and the structure diagrams of the enantiomers covered by these clauses show graphical depictions of rings. Together the three clauses for (R)-N-1-(α-naphthyl)ethylaminocarbonyl-(S)indoline-2-carboxylic acid cover 18 of the 21 successful separations on this selector in the data set. Each one excludes all the failed separations on the selector.

a)

```
separates_on(A, '(R)-N-1-(alpha-naphthyl)ethylaminocarbonyl-
                              (S)indoline-2-carboxylic_acid') :–
              has_feature(alkyl, B, A, two_or_three),
              has_feature(amine, closest, A, one_or_two).
```

Enantiomers will separate on (R)-N-1-(α-naphthyl)ethylaminocarbonyl-(S) indoline-2-carboxylic acid if they have:–
1. an alkyl chain two or three bonds away from the chiral centre.
2. at least one amine group and the closest such group to the chiral centre is one or two bonds away.

b)

```
separates_on(A, '(R)-N-1-(alpha-naphthyl)ethylaminocarbonyl-
                              (S)indoline-2-carboxylic_acid') :–
              has_feature(bg6, second_closest, A, more_than_three),
              has_feature(ring, closest, A, zero_or_one).
```

Enantiomers will separate on (R)-N-1-(α-naphthyl)ethylaminocarbonyl-(S) indoline-2-carboxylic acid if they have:–
1. at least two six-membered aromatic rings and the second closest of these to the chiral centre is more than three bonds away.
2. at least one ring and the closest ring to the chiral centre is either at the centre or one bond away from the chiral centre.

c)

```
separates_on(A,'(R)-N-1-(alpha-naphthyl)ethylaminocarbonyl
                           -(S)indoline–2-carboxylic_acid') :–
             has_feature(rg5, closest, A, one).
```

Enantiomers will separate on (R)-N-1-(α-naphthyl)indoline-2-carboxylic acid if they have at least one five-membered aliphatic ring and the closest such ring to the chiral centre is one bond away.

Fig. 2. Clauses induced by Progol from the data on attempted separations on (R)-N-1-(α-naphthyl)ethylaminocarbonyl-(S)indoline-2-carboxylic acid.

It is impossible to justify the clauses by referring to the paper from which the data was taken because the paper does not attempt to rationalise the separations that it reports. However a booklet [5] produced by a company which supplies the CSP chiral selectors does provide some justification for the clauses.

Table 1. Numbers of separations and accuracies.

CSP chiral selector	Number of Separations			Accuracy(%)	
	Successful		Failed[a]	Training	Test[b]
	In the data set	Covered by clauses			
[c]	21	18	19	93	65
[d]	19	16	21	93	58
[e]	4	4	14	100	78
[f]	3	2	16	95	84[h]
[g]	2	2	15	100	88[h]

[a] The number of failed separations in the data set is equal to the number of failed separations excluded by the clauses.

[b] Estimate obtained from a ‘leave-one-out’ cross-validation.

[c] (R)-N-1-(α-naphthyl)ethylaminocarbonyl-(S)indoline-2-carboxylic acid

[d] (R)-N-1-(α-naphthyl)ethylaminocarbonyl-(S)-tert-leucine

[e] (R)-N-1-(α-naphthyl)ethylaminocarbonyl-(S)-proline

[f] (S)-N-1-(α-naphthyl)ethylaminocarbonyl-(S)-proline

[g] (R)-N-(3, 5-dinitrobenzoyl)naphthylglycine

[h] No clauses were induced for two of the partitions for this selector; an accuracy of 0% was assigned to these partitions when estimating the test accuracy.

5 Conclusions

A scheme for data mining a relational database of published enantioseparations has been described. As far as this author is aware, this is the first paper to describe the application of Progol to enantioseparations.

The scheme was applied to published data concerning 197 attempted separations on seven CSP chiral selectors. Progol induced a set of clauses for each of five of these selectors. All of these clauses are very concise which facilitates both their interpretation and the comprehension of their coverage. The two sets of clauses that were induced from the two training sets with a significant number of positives have some chemical justification because some aspects of these clauses reflect advice given in a booklet produced by a company that supplies CSPs. The training accuracy and test accuracy for the union of these two datasets are 93% and 61% respectively.

The results suggest that the application of ILP to enantioseparations may prove fruitful and that this line of research should be pursued further.

6 Acknowledgements

The funding was provided by the EPSRC, under the remit of the Total Technology programme, and by Zeneca Pharmaceuticals. C.H.Bryant should like to thank Dr A.Srinivasan of the Oxford University Computing Laboratory and Mr R.Soloman of Phenomenex UK Ltd for their help.

References

1. Bratko, I., Muggleton, S.: Applications of Inductive Logic Programming. *Communications of the ACM* **38**(11) (1995) 65-70
2. Bryant, C.H.: Data Mining for Chemistry: the Application of Three Machine Induction Tools to a Database of Enantioseparations. Ph.D. Thesis. University of Manchester Institute of Science and Technology, UK. 1996.
3. Bryant, C.H., Adam, A.E., Taylor, D.R., Rowe, R.C.: Towards an Expert System for Enantioseparations: Induction of Rules Using Machine Learning. *Chemometrics and Intelligent Laboratory Systems* **34** (1996) 21-40
4. Bryant, C.H., Adam, A.E., Taylor, D.R., Rowe, R.C.: Using Inductive Logic Programming to Discover Knowledge Hidden in Chemical Data. *Chemometrics and Intelligent Laboratory Systems*, **36** (1997) 111-123
5. *Chirex. The Innovative Direction in Chiral Separations.* Phenomenex Ltd. UK, Macclesfield, Cheshire, UK.
6. Cleveland, T.: Pirkle-Concept Chiral Stationary Phases for the HPLC Separation of Pharmaceutical Racemates. *Journal of Liquid Chromatography* **18**(4) (1995) 649-671
7. Dzeroski, S., Schulze-Kremer, S., Heidtke, K.R., Siems: K., Wettschereck, D.: Applying ILP to Diterpene Structure Elucidation from ^{13}C NMR Spectra. *Presented at a workshop in Bari, Italy on* 2^{nd} *July 1996 entitled 'Data Mining with Inductive Logic Programming' associated with the* 13^{th} *International Conference on Machine Learning.*
8. Muggleton, S.: Inverse Entailment and Progol. *New Generation Computing* **13**(3-4) (1995) 245-286
9. Muggleton, S., Feng, C.: Efficient Induction of Logic Programs, in: Arikawa, S., Goto, S., Ohsuga, S., Yokomori, T. (eds.), Proc. 1^{st} Conf. on Algorithmic Learning Theory, Japanese Society for Artificial Intelligence, Tokyo, 1990
10. Quinlan, J.R.: Learning Logical Definitions from Relations. *Machine Learning* **5** (1990) 239-266
11. Taylor, D.R., Maher, K.: Chiral Separations by High-Performance Liquid Chromatography. *Journal of Chromatographic Science* **30** (1992) 67-85

This article was processed using the LaTeX macro package with LLNCS style