Molecular similarity considerations in the licensing of orphan drugs

Pedro Franco[1] (jpfranco73@hotmail.com), Nuria Porta (Nuria.Porta@icr.ac.uk), John D. Holliday (j.d.holliday@sheffield.ac.uk) and Peter Willett (p.willett@sheffield.ac.uk**)**

11Information School, University of Sheffield, 211 Portobello Street, Sheffield S1 4DP, UK Telephone: +44 (0) 7768191420

Abstract

The large costs associated with modern drug discovery mean that governments and regulatory bodies need to provide economic incentives to promote the development of orphan drugs, i.e., of medicinal products that are designed to treat rare disease that affect only small numbers of patients. Under EU legislation a medicine can only be authorised for treating a specific rare disease if it is not similar to other orphan drugs already authorised for that particular disease. This paper discusses the use of 2D fingerprints to calculate the Tanimoto similarity between potential and existing orphan drugs for the same disease, and presents logistic regression models correlating these computed similarities with the judgements of human experts.

Keywords: Molecular similarity; orphan drug legislation; rare diseases

Teaser: Similarity of medicines in the context of the orphan drug legislation.

Highlights:

- Authorisation and evaluation of orphan drugs.
- Similarity in the in the context of the orphan drug legislation.
- 2D fingerprints used to evaluate similarity of orphan medicines.
- Human assessment versus computer assessment.

Introduction

An orphan drug is a medicinal product intended for the treatment of a rare disease that affects only a small number of patients (less than 200,000 individuals in the USA [Orphan drug Act, Public Law 97-414, 4 January 1983 (US)] or less than five in 10,000 individuals in the EU [Regulation (EC) No. 141/2000 of the European Parliament and Council of 16 December 1999]) [1][Heemstra, H.E., PhD thesis, University of Utrecht, 2009 at http://dspace.library.uu.nl/handle/1874/37546]. This means that orphan drugs are not an immediately attractive market for the pharmaceutical industry since the limited numbers of potential patients are unlikely to yield revenues sufficient to cover the huge costs of modern-day drug discovery programmes [2, 3]. For this reason, regulatory authorities have brought forward legislation to encourage pharmaceutical companies and research groups to develop orphan drugs by providing a range of incentives: the USA's Orphan Drug Act was introduced in 1983

and the EU's Regulation (EC) No 141/2000 in 2000. These offer a similar range of incentives, with those in the EU including grants, market exclusivity, the possibility of an accelerated review, financial incentives provided by some member states and European programmes, fee reductions, free protocol scientific advice, and regulatory support [1, 4]. These measures have been highly successful, to the point that orphan drugs comprised over half of the new molecular entities approved by the US Food and Drug Administration in 2015 [5].

Market exclusivity is arguably the most important of these incentives: under the EU legislation, a pharmaceutical company that develops an orphan drug for a specific, rare therapeutic indication is given a ten-year period of market exclusivity, during which no products that are considered to be similar to that orphan drug cannot be accepted or authorised by any European regulatory competent authority. The first orphan drug approved for a certain rare disease hence has less competition than does a conventional medicinal product, which should encourage pharmaceutical companies to invest in research aimed at identifying novel medicines for such diseases.

How then should the similarity, or non-similarity, of two molecules be judged? According to article 3 (3) of the Commission Regulation (EC) No 847/2000, a similar medicinal product is defined as a medicinal product containing a similar active substance to that included in an orphan drug already authorised in the EU for the same therapeutic indication [6]. The assessment of similarity between two medicinal products (article 8 of Regulation (EC) No 141/2000) takes three criteria into account: the molecular structure; the mechanism of action; and the therapeutic indication. Two medicinal products will be considered not similar if there are significant differences under one or more of these three criteria [7]. Thus far, the European Medicines Agency (EMA), the regulatory authority that is responsible for evaluation of medicines throughout the EU, has used human judgments of similarity when assessing new medicines for rare diseases. In a previous paper, we described the first stage of a research project to provide computational support for these judgements [9]; here, we extend this work by testing further types of 2D fingerprint and describe a web application for this purpose that is based on open-source chemoinformatics software.

It should be noted that the European Commission has recently undertaken a public consultation on the concept of 'similar medicinal product' in the context of orphan drug legislation, since it is now over 15 years since the implementation of Commission Regulation (EC) No 847/2000 and the Commission aims to collect views, relevant evidence and information from stakeholders to help it develop its thinking in this area [8]. The definition of similarity in the consultation document is still quite vague and leaves considerable room for subjective interpretation, and we hence believe that the quantitative methodology described in this article provides an important contribution to the ongoing debate.

The next section describes the methods that were used by Franco et al. [9] and that form the basis for the experiments reported here. The paper continues with the results that were obtained when

these methods were used with open-source fingerprint software, and this is followed by a description of a software application that we have developed using these results. Further details of this work are reported by Franco [Franco. P., thesis, University of Sheffield, 2015 at http://etheses.whiterose.ac.uk/10148/]

Methods

The judgement as to whether a new medicine should be authorised as an orphan drug status in the EU is taken by a panel of human experts, the EMA's Committee for Medicinal Products for Human Use(CHMP). To mimic this process, Franco et al. [9] selected a set of 100 pairs of bioactive substances from the Drug Bank 3.0 database [10], and each of 143 quality experts from the EU, Japan and the USA was asked to consider each such pair and to decide whether they were, or were not, similar to each other. These binary assessments (i.e., "similar" or "not similar") were then compared with similarity values obtained by calculating the similarity between the 2D fingerprints describing the two molecules using the Tanimoto similarity coefficient [11]. The proportion of the expert assessors deciding that each pair was similar was then correlated with the computed similarity values using logistic regression. Specifically, regression models were developed to predict the probability (p) that the human assessors would decide that a pair of molecules were indeed similar given the value (x) of the Tanimoto similarity coefficient for that pair. The models are thus of the form

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = a + bx$$

where a and b are the regression coefficients. The models, which had been trained using the DrugBank data, were then shown to possess considerable predictive power when they were applied to real data that had been previously evaluated by the CHMP. Specifically, the test-set contained 100 molecule-pairs in which one molecule was an existing orphan drug for some specific rare disease and the other was a molecule that had been submitted to the CHMP for consideration for orphan drug status for that disease[1]. Given the success of this study, it was concluded that measures of similarity based on 2D fingerprints might provide a useful source of information to the CHMP during the evaluation of a new application for orphan-drug status.

The study evaluated six different types of 2D fingerprint (specifically BCI, Daylight, ECFC4, ECFP4, MDL and Unity), but all of them were proprietary in nature and thus not necessarily available to any organisation or individual wishing to use our methods. The work described here was undertaken to overcome this limitation of our previous study, and involved using fingerprints available in the open-source CDK and RDKIT systems (at https://sourceforge.net/projects/cdk/ and http://www.rdkit.org/RDKit_Docs.current.pdf, respectively). Specifically, the following twelve types

[1] It is not possible to include the CHMP molecules on grounds of commercial confidentiality, but the DrugBank dataset is available as supplementary information for the paper by Franco et al. at https://jcheminf.springeropen.com/articles/10.1186/1758-2946-6-5

of fingerprint were generated, as implemented in the KNIME pipelined data analysis system (at http://www.knime.org): CDK Extended, CDK Standard, Estate, PubChem, MACCS, Morgan, Feat Morgan, Atom Pair, Torsion, RDKit, Avalon, and Layers. These fingerprints cover all of the many types used in current chemoinformatics systems (circular, hashed, fragment dictionary, and topological patterns) [12]. Logistic regression models were developed for each type of fingerprint, correlating the Tanimoto similarity for a pair of molecules with the probability of the human experts deciding that that pair was similar; the training dataset was the 100 pairs of molecules in the DrugBank dataset; and the resulting models were then validated using the 100 pairs of molecules in the CHMP test dataset.

Results and discussion

The results obtained with the DrugBank training set will be illustrated using the CDK Extended fingerprint, a hashed fingerprint that is similar in concept to Daylight fingerprints and that encodes all linear paths up to eight atoms in length; in addition, the fingerprint contains bits detailing the numbers of fused and unfused rings.

Figure 1 plots the proportion of the expert assessors who judged a molecule-pair as being similar (Y-axis) against the computed fingerprint similarity score for that molecule-pair (X-axis). There is a well-marked separation of the similar pairs (marked in green) and the non-similar pairs (marked in blue), with the smaller Tanimoto values dominating the lower left portion of the curve (i.e., only a small proportion of the experts judged that a pair should be considered similar) and the larger Tanimoto values the upper right portion (i.e., most of the experts judged that a pair should be considered similar). The solid line in this figure represents the estimated probability of being similar as predicted by the logistic regression model, together with the 95% confidence limits for this prediction (the dotted lines). The values of a and b in the regression model for this fingerprint were -16.761 and 2.881 respectively, with these values being statistically significant ($p = 0.0012$) and with the Nagelkerke R^2 value of 0.917 indicating a good fit of the model to the data. The CDK Extended fingerprint gave the highest Nagelkerke value for all of the twelve fingerprints tested here; the lowest value of 0.690 was obtained with the EState fingerprint (which encodes structural keys but is, at 79 bits, by far the shortest of all the fingerprints).

If a logistic regression model is to be used predictively then one must be able to identify an appropriate threshold similarity for deciding that two molecules should indeed be considered as similar. Let t denote the threshold similarity such that a pair of molecules are predicted to be similar if their computed similarity is $\geq$ t and predicted not to be similar if $<$ t. The predictions resulting from use of a particular value for t can then be compared with the judgements of the 143 experts, thus allowing the calculation of the numbers of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) where, e.g., TP is the number of cases where the majority of the experts judged two molecules to form a similar molecule-pair and where those two molecules had a computed

similarity $\geq$ t. Knowledge of these four values then allows the plotting of a receiver operating characteristic (ROC) curve, which demonstrates the relationship between the specificity (TN/(TN+FP)) and the sensitivity (TP/(TP+FN)) as the threshold, t, is systematically varied. The resulting area under the ROC curve, or AUC, was 0.992, this very high value indicating the ability of the CDK Extended fingerprint to discriminate between similar and non-similar pairs of molecules. This was the largest AUC value obtained across the twelve fingerprints, with the EState fingerprint, as with the Nagelkerke statistic, giving the lowest value (of 0.929).

Several other measures based on these four variables (i.e., TP, TN, FP and FN) are available in the literature to measure the effectiveness of a predictive system. Those computed here were the precision, the accuracy, the F index, the Youden index, and Matthew's correlation coefficient, as described by Franco et al. [9]. These five statistics were computed as t was varied, thus enabling us to identify that threshold similarity value that gave the best predictions, where this optimal value was taken to be that which gave the largest values for the precision, the accuracy, the F index, the Youden index and the Matthews coefficient whilst at the same time providing acceptable values for the sensitivity and the specificity. For the CDK Extended fingerprint, a value of 0.610 for t was found to be the best, and this was then used for the analysis of the CHMP test dataset. An entirely comparable process was used to determine the optimal similarity threshold for each of the other eleven types of fingerprint. In five cases (MACCS, Morgan, Feat Morgan, Torsion and Layers), two different values of t were found to yield comparable results on the training dataset so just a single value (specifically, those obtained with the larger of the two t values) is quoted in the results in Table 1 that are discussed below.

The regression models developed on the DrugBank training dataset were then applied to the CHMP test dataset, as detailed in Table 1. Each row of the table is associated with one type of fingerprint and lists the a and b coefficients from the regression model, the Nagelkerke R^2 and AUC values for that model, the threshold similarity t as derived above, and finally the number of test cases that were predicted correctly, where a correct prediction is a pair of molecules for which the prediction from the model, (i.e., similar or not similar) mirrors that arrived at by experts comprising the CHMP panel. For comparison, the bottom row of the table contains the corresponding data for the BCI (Barnard Chemical Information) fingerprint that was found to give the best overall level of performance in our previous comparison of six different types of proprietary fingerprint. Inspection of Table 1 shows that all but three of the fingerprints (Estate, Atom Pair and RDDKit) were able to predict successfully the similarity or non-similarity of 95 or more of the pairs of molecules in the test dataset, with the Morgan fingerprint resulting in only a single incorrect prediction (despite it achieving only moderate R^2 and AUC values in the training stage). The incorrect prediction was for a pair of active substances in which both had complex structures and molecular weights in excess of 500. It was not just the Morgan fingerprint that failed here, since all of the models that were developed failed to make a correct prediction for this particular case.

Data fusion, or consensus scoring, has been found to be of considerable value for similarity-based virtual screening [13] and we have adopted this approach to the present context. Specifically, assume that each of several different fingerprints is used to make a prediction for a pair of molecules: then the final prediction is that given by the majority of the fingerprints. It was found that 99 correct predictions were obtained using both all of the CDK fingerprints (i.e., four or more of them giving the same prediction) or all twelve fingerprints (i.e., seven or more of them giving the same prediction). The incorrect predictions were again for the pair of molecules discussed in the previous paragraph. Since different fingerprints capture different characteristics of molecular structure and since the open-source nature of CDK and RDKit provides a range of such fingerprints, a consensus approach such as this would seem to be appropriate for future applications of our approach.

One such application that has been developed is a KNIME-based system that is now being used to support assessors throughout the EU who assist CHMP in coming to a decision as to whether a medicine for a rare disease should be approved.. The assessors do this by producing an evaluation report on similarity in which they comment on the extent to which the new active substance is indeed sufficiently distinct from existing orphan drugs for that indication. The application takes as input the SMILES linear notations of the two (or more) molecules that are to be compared; generates the required fingerprints (initially CDK Extended, CDK Standard, Avalon and Layers as these were the ones where the computed similarity values correlated most strongly with the BCI fingerprint that performed best in our previous study); computes the Tanimoto similarity coefficient using each of these fingerprints; and outputs an Excel spreadsheet containing the coefficient values and the corresponding probabilities of being judged similar as calculated using the appropriate regression models from Table 1. The assessor can then use this spreadsheet when coming to a conclusion as to the novelty or otherwise of the submitted molecule, the report that is submitted to the EMA being based not just on the fingerprint-based structural similarity but also on the reported therapeutic indication and mode of action. The sets of assessor reports provide an important input to the CHMP's final decision as to whether or not to assign orphan drug status to a molecule that has been submitted for consideration. In addition, of course, pharmaceutical companies can use the data in Table 1 to assist them prior to submission of a new molecule to the EMA for consideration for marketing authorisation.

In addition to the 2D fingerprints discussed here, comparable experiments were also conducted using two other types of molecular representation [P. Franco, PhD these, University of Sheffield, 2015 at http://etheses.whiterose.ac.uk/10148/]. In the first of these, molecules were characterised by 23 computed molecular properties (such as molecular weight, logP, molar refractivity, numbers of rotatable bonds etc.); and in the second by four further proprietary fingerprints produced by the Chemical Computing Group as part of their Molecular Operating Environment system (at https://www.chemcomp.com/): two of these – Typed Graph Distances and Typed Graph Triangles – described the 2D shape of a molecule while the other two – Typed Atom Distances and Typed Atom Triangles – described the 3D shape of a molecule. None of these

alternative types of molecular representation were found to perform as well as the 2D fingerprints used here. There are, of course, many other ways in which structural similarity can be computed [14, 15]. A fingerprint is a very simple, indeed crude, description of a molecule's 2D structure, and more complex similarity measures have hence been described that are based on the maximum common subgraph (or MCS), i.e., the largest subgraph common to the graphs describing two 2D molecular structures [16, 17]. The MCS provides an intuitive measure of similarity but does have three limitations. First, an MCS can be defined in multiple ways: it may be based upon the numbers of matching atoms and/or bonds; and it can be connected, be disconnected, or be disconnected subject to the individual components being of at least some threshold size. Next, a similarity measure can be based upon just the MCS (or MCSs if there is more than one) or upon a weighted combination of the MCS and of the smaller maximal common subgraphs. Finally, MCS detection is computationally demanding, and many of the MCS algorithms that have been published are non-deterministic or approximate in nature. In view of the very high level of performance achieved in our experiments, we believe that 2D fingerprints provide the most appropriate way of investigating the similarity relationships pertinent to the registration of orphan drugs.

Conclusions

Measures of structural similarity based on 2D chemical fingerprints are widely used in medicinal chemistry (and in chemoinformatics more generally) for applications such as database clustering, diversity analysis and ligand-based virtual screening.

In our work we have shown how this very common type of structural representation can also be applied to the licensing of medicines for rare diseases by regulatory authorities. Experiments with a range of fingerprints that can be generated using widely available, open-source software show that they provide measures of Tanimoto-based similarity that correlate well with expert assessments of structural similarity; and that logistic regression models based on these similarities and human assessments mirror very closely the final pronouncements of the EMA, the EU's regulatory authority for the licensing of orphan drugs. A software application based on our findings provides a simple tool to support the work of EU experts who contribute to the final recommendations made to the EMA by CHMP.

Acknowledgements:

We thank Pedro Oliveira for advice on web applications and Dr. Jean-Louis Robert for helpful discussions on similarity.

Bibliography

1. Franco, P. (2013) Orphan drugs: the regulatory environment. Drug Discov. Today, 18, 163-172.
2. DiMasi J.A. et al. (2016) Innovation in the pharmaceutical industry: new estimates of R&D costs. J. Health Econ., 47, 20-33
3. Morgan, S. et al. (2011) The cost of drug development: a systematic review. Health Policy, 100, 4-17.
4. European Commission (2003) Communication from the Commission on Regulation (EC) No 141/2000 of the European Parliament and of the Council on orphan medicinal products – (2003/C 178/02) (29 July 2003).
5. Kinch, M.S (2016) 2015 in review: FDA approval of new drugs. Drug Discov. Today, 21, 1046-1050.
6. European Parliament and Council (2000) Regulation (EC) No. 847/2000 of the European Parliament and Council (27 April 2000).
7. European Commission (2008) Communication from the Commission: Guideline on aspects of the application of Article 8(1) and 3 of Regulation (EC) No 141/2000: Assessing similarity of medicinal products vesus authorized orphan medicinal products benefiting from market exclusivity and applying derogations from that market exclusivity – C (2008) 4077 (19 September 2008).
8. European Commission (2016) Consultation document on the Concept of "similar medicinal product" in the context of the Oprhan Legislation: Adptation to Technical Progress – (29 July 2016)
9. Franco, P. et al. (2014) The use of 2D fingerprint methods to support the assessment of structural similarity in orphan drug legislation. J. Cheminf., 6(1), 5.
10. Wishart D.S. et al. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res., 36, D901-9066.
11. Willett, P. (2006) Similarity-based virtual screening using 2D fingerprints. Drug Discov. Today, 11, 1046-1053
12. Leach A.R. and Gillet V.J. (2007) An Introduction to Chemoinformatics, 2nd edition, Springer.
13. Willett, P. (2013). Combination of similarity rankings using data fusion. J. Chem. Inf. Model., 53, 1-10
14. Willett, P. (2009) Similarity methods in chemoinformatics. Ann. Rev. Inf. Sci. Technol., 43, 3-71.
15. Maggiora, G. et al. (2014) Molecular similarity in medicinal chemistry. J. Med. Chem., 57, 3186-3204.
16. Raymond, J.W. and Willett, P. (2002) Maximum common subgraph isomorphism algorithms for the matching of chemical structures. J, Comput.-Aided Mol. Design, 16, 521-533.
17. Ehrlich, H.-C. and Rarey, M. (2011) Maximum common subgraph isomorphism algorithms and their applications in molecular science: a review. WIRES Comput. Mol. Sci., 1, 68–79

Figures and Tables

Figure 1. Correlation between the probability of a pair of molecules being similar based on human judgements and the Tanimoto coefficient using the CDK Extended fingerprint

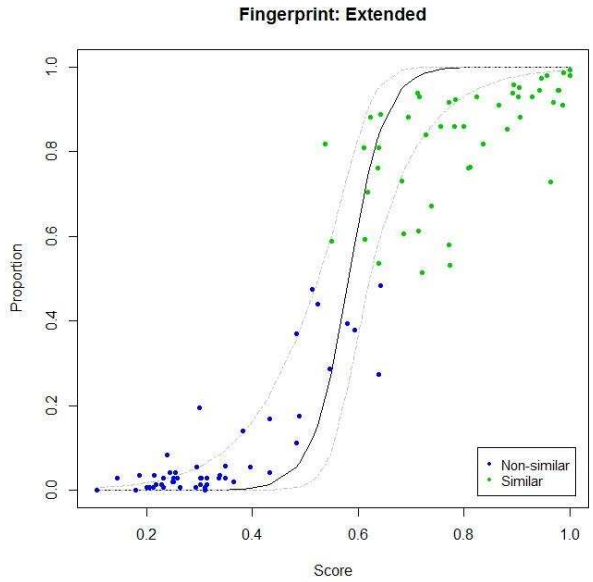

Table 1. Logistic regression models for different 2D fingerprints

Fingerprint	a	b	R^2	AUC	t	Predictions
CDK Extended	-16.761	2.881	0.917	0.992	0.610	96
CDK Standard	-13.971	2.446	0.905	0.989	0.587	96
EState	-9.642	1.334	0.690	0.929	0.714	91
PubChem	-12.847	1.717	0.772	0.957	0.776	95
MACCS	-9.071	1.398	0.811	0.974	0.750	95
Morgan	-8.204	2.238	0.871	0.982	0.400	99
Feat Morgan	-9.041	1.822	0.876	0.987	0.519	96
Atom Pair	-11.790	2.492	0.879	0.984	0.459	80
Torsion	-6.533	1.877	0.882	0.986	0.370	98
RDKit	-8.249	1.333	0.788	0.964	0.671	73
Avalon	-8.447	1.536	0.881	0.987	0.625	95
Layers	-22.314	3.031	0.877	0.985	0.761	95
BCI	−12.758	2.128	0.906	0.990	0.606	97