International Journal of Pharmaceutics 514 (2016) 335–340

Contents lists available at ScienceDirect

International Journal of Pharmaceutics

journal homepage: www.elsevier.com/locate/ijpharm

Use of similarity scoring in the development of oral solid dosage forms

Ana P. Ferreira[a,*], Dolapo Olusanmi[b], Omar Sprockel[b], Admassu Abebe[b], Faranak Nikfar[b], Mike Tobyn[a]

[a] Drug Product Science & Technology, Bristol-Myers Squibb, Reeds Lane, Moreton, Wirral, CH46 1QW, UK
[b] Drug Product Science & Technology, Bristol-Myers Squibb, 1 Squibb Drive, New Brunswick NJ08903, USA

ARTICLE INFO

Article history:
Received 29 June 2015
Received in revised form 18 September 2015
Accepted 19 September 2015
Available online 28 September 2015

Keywords:
Solid oral dosage form
Quality by design
Robustness
Similarity scoring
Surrogate
Knowledge management

ABSTRACT

In the oral solid dosage form space, material physical properties have a strong impact on the behaviour of the formulation during processing. The ability to identify materials with similar characteristics (and thus expected to exhibit similar behaviour) within the company's portfolio can help accelerate drug development by enabling early assessment and prediction of potential challenges associated with the powder properties of a new active pharmaceutical ingredient. Such developments will aid the production of robust dosage forms, in an efficient manner.

Similarity scoring metrics are widely used in a number of scientific fields. This study proposes a practical implementation of this methodology within pharmaceutical development. The developed similarity metrics is based on the Mahalanobis distance. Scanning electron microscopy was used to confirm morphological similarity between the reference material and the closest matches identified by the metrics proposed. The results show that the metrics proposed are able to successfully identify material with similar physical properties.

1. Introduction

The performance of an oral solid dosage (OSD) formulation is heavily influenced by the physical properties of the active ingredient and excipients that compose it (Hlinak et al., 2006; Leane et al., 2015). The key challenges associated with manufacturing of OSD are ensuring that the blends have good flow and are able to form strong compacts, containing the right amount of drug, which then disintegrate and release the drug into the body at the predetermined rate.

During the early stages of drug development often only small amounts of active pharmaceutical ingredient (API) are available to conduct formulation and process development, and identify the challenges associated with the powder properties of the API. Effective extraction of the experience accumulated by a pharmaceutical company over the years (i.e. knowledge management) will bring immense advantage to the scientist tasked with the development of a new formulation. The ability to anticipate risks and identify the optimal formulation would yield a drug product that is fit for the intended purpose. It also will help in the identification and implementation of the most appropriate, i.e. robust, route of manufacture.

The question is then how to identify, from the body of knowledge gathered over the years, the relevant data for the new compound under development. This approach has been proposed as identification of a "surrogate" API (Hancock, 2010). The identification of the appropriate surrogate for a new compound is, in essence, a classification problem and one which can be solved through the use of similarity metrics.

Similarity metrics have been used for several decades in the context of cluster analysis. Cluster analysis is the term applied to a number of methods that study the existence of groups within a set of data (Bratchell, 1989). These methods require the quantification of similarity or dissimilarity between observations in the data set, which for the most part is based on the concept of distance between objects (Bratchell, 1989). Several metrics have been proposed, see e.g. (Brereton, 2003b), and a list of those most commonly used is provided in Table 1. The term *similarity metrics* has been used in over 2000 papers (Elsevier, 2014a) and the term *similarity scoring* appears in 82 publications (Elsevier, 2014b). The areas with a higher number of applications reported in literature are computer science; engineering; mathematics; medicine; biochemistry, genetics and molecular biology; physics and astronomy (Elsevier, 2014a).

* Corresponding author. +44 151 552 1522.
E-mail address: ana.ferreira@bms.com (A.P. Ferreira).

http://dx.doi.org/10.1016/j.ijpharm.2015.09.046
0378-5173/

Table 1
Common similarity metrics.

	Definition	Notes
Correlation coefficient (Pearson)	$R = \frac{\sum_{j=1}^{J}(x_{kj} - \bar{x}_k) \times (x_{lj} - \bar{x}_l)}{\sqrt{\sum_{j=1}^{J}(x_{kj} - \bar{x}_k)^2 \times \sum_{j=1}^{J}(x_{lj} - \bar{x}_l)^2}}$	k,l – observations in a data set with J variables x_{ij} is the j^{th} measurement on sample i C is the variance-covariance matrix of the variables
Euclidean distance	$d_{kl} = \sqrt{\sum_{j=1}^{J}(x_{kj} - x_{lj})^2}$ or $d_{kl} = \sqrt{(x_k - x_l) \times (x_k - x_l)^T}$	
Manhattan distance	$d_{kl} = \sum_{j=1}^{J}\lvert x_{kj} - x_{lj}\rvert$	
Mahalanobis distance	$d_{kl} = \sqrt{(x_k - x_l) \times C^{-1} \times (x_k - x_l)^T}$	

Within the pharmaceutical industry, similarity metrics have been proposed for use in in-silico prediction of the activity of large libraries of compounds with the purpose of identifying the most promising leads for further experiments (Kristensen et al., 2013; Willett, 2006) and screening multiple candidate compounds against off-target effects to elucidate possible liabilities (Wendt et al., 2011). This type of metrics have also been applied in automated polymorph screening and salt selection (Ivanisevic et al., 2005) and to review data from bioequivalence studies (Polli and McLean, 2001).

Similarity scoring also has diverse applications outside the scientific disciplines. In sports, algorithms have been used to estimate statistics for players and teams performance over the next season: baseball – PECOTA (Silver, 2003), American football – KUBIAK (Schatz, 2008), basketball – SCHOENE (Pelton, 2008) and ice-hockey – VUKOTA (Awad, 2009). Another area where this approach is used extensively is in E-commerce with companies employing similarity metrics to enable targeted marketing approaches which provide personalised recommendations to customers based on order history or searched-for items (Greg et al., 2003).

In this paper, we propose the application of similarity metrics to identify active pharmaceutical ingredients with similar physical properties with the aim of facilitating development of oral solid dosage forms. Materials with similar physical characteristics should exhibit similar behaviour during manufacturing (Hancock, 2010); therefore, similarity metrics applied to a data set with physical properties from a number of compounds will enable the identification of materials likely to behave similarly during drug product manufacturing. The ability to identify similar materials can be used during oral solid dosage development to support early risk assessment activities (e.g. leveraging lessons learned from formulation challenges encountered for materials with similar properties will highlight the potential pitfalls for the new asset).

The similarity scoring metrics proposed in this manuscript is based on the Mahalanobis distance (De Maesschalck et al., 2000) combined with the application of principal component analysis (PCA) (Esbensen and Geladi, 2009). The example provided will illustrate how this metrics was used to identify materials which are the most similar to a particular batch.

2. Material and methods

2.1. Sample characterization

Particle size and shape distributions were measured using a Morphologi G3 particle characterization system (Malvern Instruments Limited, Malvern, UK). Samples were suspended in octane and dispersed using gentle shaking. The suspensions were pipetted onto microscope slides and left to dry. The dried slides were then analysed using method and lens configurations appropriate for each material and the final image data morphologically filtered using a range of standard filters, such as solidity and convexity, to remove partially imaged and/or overlapping particles. For shape analysis an additional pixel number filter was applied removing all particles consisting of less than 100 pixels. Additional details can be found in (Gamble et al., 2011).

The specific surface area (SSA) was determined using the BET technique with a Gemini 2390a surface area analyser (Micromeritics, USA) and nitrogen as the adsorbate. About 0.5–1 g of material was analysed per batch. To ensure accurate surface area determination, samples were degassed prior to analysis for 24 h at 50 °C using nitrogen gas to remove residual moisture adsorbed on the surface of the particles which might affect the accuracy of the results. These conditions were identified as suitable for these samples during method development. During analysis, the samples were equilibrated for 10 s followed by an evacuation rate of 66.6 kPa/min. Multi-point B.E.T. measurements were taken in the range 0.05–0.3 p/p0. Two samples were analysed for each batch and the average determined.

Samples were imaged using a scanning electron microscope Neoscope JCM-500 (Jeol Inc., MA, USA). Samples were sputter coated prior to imaging using a JFC-1300 auto fine coater (Jeol Inc, MA, USA).

2.2. Data set

The internal database of material properties was queried to obtain data for all samples for which particle size, shape and SSA measurements had been carried out. The search retrieved data for 107 lots from 28 different compounds. The search results were exported from the database in excel format. The data set used in this analysis has 13 parameters: the 10, 50, and 90 percentiles of the size and shape distributions (both the number based and volume based distributions of the equivalent spherical diameter were included) (Alan, 2008) and the specific surface area. The amount of material required to measure particle size, shape and SSA on a sample of material is less than 2 g.

2.3. Choice and implementation of similarity metrics

A good indication of the similarity between two objects is a measure of the distance between them. The most commonly used