THE SIGNAL: STATISTICAL ASPECTS, NORMALISATION, ELEMENTARY ANALYSIS

Samuel WIECZOREK

4.1. INTRODUCTION

The elementary analysis of raw data coming from automated pharmacological screening (i.e. the bioactivity signals) aims to identify bioactive molecules (called candidate hits) that will then be subjected to more in-depth testing. This selection is made by setting a bioactivity threshold and the interesting molecules are therefore identified purely on the basis of the bioactivity signal. This measure represents the most concise information about the bioactivity of compounds in a chemical library and is as such particularly precious.

During automated screening, the bioactivity signals are characterised by variability and uncertainty due to measurement errors (fig. 4.1), which may have a biological, chemical or technological origin. These errors give rise to false-positives (molecules wrongly identified as bioactive) as well as false-negatives (molecules identified as bio-inactive despite having actual bioactivity). These phenomena degrade the quality of the selection of bioactive molecules.

Fig. 4.1 - A threshold for the measured signal permits selecting the molecules of interest

(a) Ideal case - Measurements without errors: the signals and the bioactivity threshold are precise. (b) Real case - Measurements marred by errors: the signals as well as the bioactivity threshold are imprecise.
The validity of the conclusions drawn from the elementary analysis depends on the quality of the underlying raw data. Would pre-processing of the raw signals help to improve the precision of the information and to limit the influence of errors on the results?

4.2. Normalisation of the signals based on controls

The variability within the data arising from screening complicates the identification of bioactive molecules. Considering the whole set of data for a given screening, the selection is carried out by using a cut-off for the raw signal, which is not always comparable from one plate to another. To overcome this difficulty, the traditional approach of normalisation (by the percentage of inhibition), based on the means of the control values for bioactivity and bio-inactivity, functions correctly and remains widely used. If the side effects are not too widespread and if the controls are inspected for discrepancies and aberrant values, then normalisation by the percentage of inhibition is often valid (BRIDEAU et al., 2003).

4.2.1. Normalisation by the percentage inhibition

The Percentage of Inhibition (PI) scales the raw bioactivity signal to a value lying between 0 and 1 (and multiplied by 100 to put it on a percentage scale). For a plate $p$, the percentage inhibition $PI_p^i$ of the signal measured in a well with index $i$ represents its relative distance from the mean of a set of control bioactivity values. Let $I^{act}$ and $I^{inact}$ be the respective means of a set of controls for bioactivity and bio-inactivity and $I_p^i$, the signal from a molecule measured in a well with index $i$ in a plate $p$, the normalised signal is defined as follows:

$$PI_p^i = \frac{I_p^i - I^{inact}}{I^{act} - I^{inact}}$$  \hspace{1cm} (eq. 4.1)

The normalised signal is interpreted thus: the closer the raw signal measured is to the mean of the controls used for bio-inactivity, the more the percentage inhibition approaches 0; conversely, the closer the signal approaches the mean of the controls used for bioactivity, the more the normalised signal value tends towards unity.

Note that it is entirely possible to observe molecules for which the raw signal exceeds that of the controls (percentage inhibition $< 0$ and $> 100$).

4.2.2. Normalisation resolution

The normalisation presented in the preceding section is based on a set of controls. This particular set, termed the normalisation window (fig. 4.2), defines the controls for bioactivity and bio-inactivity whose means permit the calculation of the percentage inhibition (eq. 4.1). The width of this window, i.e. the number of con-