Cellular Fingerprints: A Novel Concept for the Integration of Experimental Data and Compound-Target-Pathway Relations

(Extended Abstract)

Stefan Gunther, Stefanie Neumann, Jessica Ahmed, and Robert Preissner

Structural Bioinformatics Group, Institute of Molecular Biology and Bioinformatics, Charité, Universitätsmedizin Berlin, Arnimallee 22, 14195 Berlin, Germany
http://bioinformatics.charite.de/

Extended Abstract

The pharmaceutical industry is hunting for high-affinity inhibitors of medical targets, but most of them fail in clinical trials because of severe side effects. On the other hand, there is a growing knowledge about multiple targets and their role in various signalling pathways. Therefore, the integration of experimental data, literature knowledge about drugs, targets, their metabolism, ontology, and related pathways is an important task to achieve better understanding of drug mechanisms on a systems biological level.

To this end we have compiled a database allowing complex queries that project various types of information onto 2,500 WHO-classified drugs [1]. Currently, the database contains about 3,000 target proteins that are annotated by more than 8,000 literature-based and manually curated drug-target relations. The use of the Anatomical-Therapeutical-Chemical drug classification (ATC-code) [2] enables easy access to medical indications or diseases and links phenotypic data to biological processes on a molecular level. Integrated Gene Ontology (GO) [3] enables filtering of proteins associated to particular molecular functions or cellular components. Similarity searching for drugs or targets is implemented via structural fingerprints [4] and FASTA-alignments. Structural fingerprints are bit-vectors encoding for the chemical and topological features of drugs and drug-like compounds. Their similarity can be described using the Tanimoto-coefficient ($T$), considering the concordant and unequal bits of two structural fingerprints [5].

$$T = \frac{N_{ab}}{N_a + N_b - N_{ab}}.$$  

$N_a$ is the number of bits set to 1 in compound $a$, $N_b$ is the number of bits set to 1 in compound $b$ and $N_{ab}$ is the number of bits common to both, compound $a$ and $b$.

For a more comprehensive approach, the public data from the National Cancer Institute on expression and cell effects related to 50,000 compounds are a valuable resource [6]. Besides the expression of all 60 cell types in their basic state, data on the changed
expression after application of cancer drugs with known mechanism of action are deposited. Moreover, data on mutations and properties of the cell panel of the NCI exist [7], which will be useful for the modelling of differences between the cell lines. However, NCI-compounds, which show low cell type specificity were excluded. The remaining data were normalised by z-normalisation to achieve a comparable level of data and to emphasize the differences in the cell specificity. In a next step, the effect on 60 cell lines was translated for each compound into a bit-vector of length 4,800. Thus, this cellular fingerprint describes the unique pattern of effects of a chemical compound on the NCI60 cell panel comparable to the structural fingerprint describing the chemical properties of a compound. Since the average number of cell lines that respond specifically to a single compound is relative low compared to the non-specific cell reactions (about 5/60), an asymmetric distance is an appropriate measure. Again, for fast similarity searching, the Tanimoto-coefficient was used.

The exciting question is whether we will find a correlation between structural and cellular similarity. If the structural similarity between two compounds correlates well with the profile of targets addressed by these compounds, this should be the case. On the other hand, it is known that even for compounds with a distinct structural similarity (Tanimoto > 85%) only one third exhibits similar effects in the same experimental assay [8]. To answer the question for correlation between structural and cellular similarity, a clustering of the compounds with similar cellular effects regarding their structural similarity was performed (Fig. 1). In general, the vast majority of compounds occurs in two or three large clusters, where each cluster represents one structural scaffold. Preliminary analysis shows that one cluster of compounds can be found addressing the same target (Fig. 2) or compounds that address another target in the same pathway. This shows that the cellular fingerprint is indicative of target specificity without structural similarity bias. Examples for coincidence of structural and cellular fingerprints are presented in Fig. 3.

![Fig. 1. Clustering of compounds with similar cellular fingerprint according to their structural similarity](image)

The presented method supports the following goals:

- Prediction of targets and mechanism of action
  Similarity of structural fingerprints allows hypotheses about similar cellular fingerprints and vice versa. Experimental data on a sufficient number of cell lines allow conclusions about the mechanism of the compound.