BMC Bioinformatics

METHODOLOGY ARTICLE

Open Access

Predicting biological pathways of chemical compounds with a profile-inspired approach

Javier Lopez-Ibañez, Florencio Pazos and Monica Chagoyen*

*Correspondence: monica.chagoyen@cnb.csic.es
Computational Systems Biology Group, National Center for Biotecnology (CNB-CSIC), Darwin 3, 28049 Madrid, Spain

Abstract

Background: Assignment of chemical compounds to biological pathways is a crucial step to understand the relationship between the chemical repertory of an organism and its biology. Protein sequence profiles are very successful in capturing the main structural and functional features of a protein family, and can be used to assign new members to it based on matching of their sequences against these profiles. In this work, we extend this idea to chemical compounds, constructing a profile-inspired model for a set of related metabolites (those in the same biological pathway), based on a fragment-based vectorial representation of their chemical structures.

Results: We use this representation to predict the biological pathway of a chemical compound with good overall accuracy (AUC 0.74–0.90 depending on the database tested), and analyzed some factors that affect performance. The approach, which is compared with equivalent methods, can in addition detect those molecular fragments characteristic of a pathway.

Conclusions: The method is available as a graphical interactive web server http://csbg.cnb.csic.es/iFragMent.

Background

Studying the roles of chemical compounds in a cellular context is fundamental for understanding living systems at the molecular level [1]. This can be achieved with experimental and computational approaches. Among the last, of special interest is the analysis of relations between the structure of a chemical compound and its biological role.

Knowledge on biological pathways is still incomplete, so pathway databases are continuously updated, both by adding new pathways, as well as molecular components to existing pathways. Predicting the biological role of a chemical compound, namely the pathway(s) it is involved in, from its chemical structure would be valuable not only in the assignment of new compounds to known biological pathways, but also in other applications like the functional interpretation of metabolomics experiments or the prediction of possible biological roles of drugs. A number of studies aimed to predict the biological pathway of a compound from its chemical structure alone. Most biological pathways contain a reduced number of metabolites, what could be a problem for machine learning approaches. This is one of the reasons

why most studies have tried to predict general pathway classes [2–7] (e.g. carbohydrate metabolism). While predicting the general pathway class of a compound would be valuable in some cases, most applications will require the prediction of specific pathways (e.g. glycolysis).

Two previous studies, in addition to predict general pathway classes, have also tried to predict specific pathways. In the context of a wider analysis of human metabolic pathways and their mapping in the chemical space, Macchiarulo et al. [8], using a machine learning algorithm (random forests), assigned compounds to 52 human metabolic pathways defined in KEGG [9]. More recently, Hamdalla et al. [10], proposed a family of related approaches based on a ranking algorithm and pair-wise substructure matching that were tested in 137 KEGG metabolic pathways. Their best performing approach (implemented as a software package, TrackSM) followed a two-step classification: in the first step, TrackSM predicts the pathway class of a compound, and in a second step it identifies the specific pathway from that class.

In this work we propose a method, iFragMent, for the prediction of compounds involved in specific pathways (not pathway classes), inspired by sequence profile approaches. Sequence profiles have been widely used in the study of DNA and protein sequences, and they are behind most modern methodologies for obtaining information from these biological polymers. Sequence profiles are formal models that capture the main characteristics of a set of related sequences. They are built from a multiple sequence alignment of these related sequences using different approaches, from simple "position specific scoring matrices" (PSSMs) to complex statistical models such as "hidden Markov models" (HMMs) [11]. Once built, these profiles allow assigning new members to the family (hence predicting their function if it is unknown), detect functionally important residues (e.g. conserved positions) or define domains, among other things. These profile-based approaches have been designed taking into account the polymeric nature of DNA and proteins and their underlying evolutionary relationships.

We explore the possibility of using a conceptually (not methodologically) related strategy for studying chemical compounds in the context of biological pathways. We take into account the chemical composition (in the form of chemical fragments) of the whole set of compounds participating in a pathway (e.g. metabolic, regulatory and signalling networks). We calculate the enrichment of chemical fragments in the pathway, and use this information to score new compounds, given the presence of the pathway-enriched fragments in their structure. We evaluate our method in its ability to predict the correct pathway of a chemical compound, using 861 pathways defined in four databases: KEGG [9], Reactome [12], SMPDB [13] and enviPath [14].

Our results show that the method proposed predicts biological pathways for chemical compounds with global AUCs (Area Under the Curve) ranging from 0.74 to 0.90 depending of the database considered. We compare our method to previous aproaches [8, 10] and to a k-nearest neighbor approach based on pair-wise structural similarities. In addition to the predicted pathway, our approach reports associated p values and detects the chemical substructures responsible for a compound-pathway assignment. The method is implemented as a web server http://csbg.cnb.csic.es/iFragMent and code is available at https://github.com/jlopez-ibanez/iFragMent.

Results

Overall performance

We tested the ability of our method to predict a compound's biological pathway(s) as defined in four databases: KEGG pathway, Reactome, SMPDB and enviPath, which represent metabolic, signaling, and biodegradative routes among others. We use only those pathways with at least 10 distinct compounds.

To explore the largest number of chemical substructures, while at the same time not imposing any a priori chemical knowledge, we represented compounds as binary fragment vectors (see "Methods" for details).

We propose a method based on the probabilities of observing the presence of structural fragments in the compounds of a pathway just by chance. A compound having fragments enriched in a given pathway is a good candidate to participate in that pathway.

We evaluated the performance of the method using a tenfold cross-validation approach (see Table 1 for overall performance results). Our method is a multi-label and multi-class approach, as it allows to assign several pathways to a compound (by taking the top-n predictions, see "Methods"). This allowed measuring performance by building ROC curves and calculate their AUC value (area under the curve), as we take as predictions from the top-1 to the full set of predictable pathways.

For all datasets we obtained good results (i.e. AUC > > 0.50). We obtained better overall results for enviPath (65 biodegradation pathways, 0.90 AUC), and KEGG (214 pathways, 0.88 AUC) than for SMPDB (333 pathways, 0.74 AUC) and Reactome (249 pathways, 0.74 AUC) databases. Neither the number of compounds nor the number of pathways was found to be related to these global performances. More details of each database can be found in Additional file 1: Table S1.

Individual pathway performance

We found that results of individual pathways varied from random to excellent performance in all databases except enviPath (with AUC > 0.7 for all pathways) (Fig. 1a).

To compare our profile-inspired approach with a pair-wise one, we implemented a k-nearest neighbour classifier (k-NN) using the same vectorial (fingerprint) representation of compounds (see "Methods"). We calculate structural similarity with the Tanimoto coefficient, widely used in chemoinformatic approaches for structural comparisons of fingerprint representations. Our method attained higher AUCs in a tenfold cross validation test than k-NN in the four databases tested (Fig. 1). Details about each individual pathway performance using both methods are provided in Additional file 2: Tables S2–S5.

Table 1 Overall performance (AUC) of the method

Database	AUC
KEGG	0.88
enviPath	0.90
SMPDB	0.74
Reactome	0.74

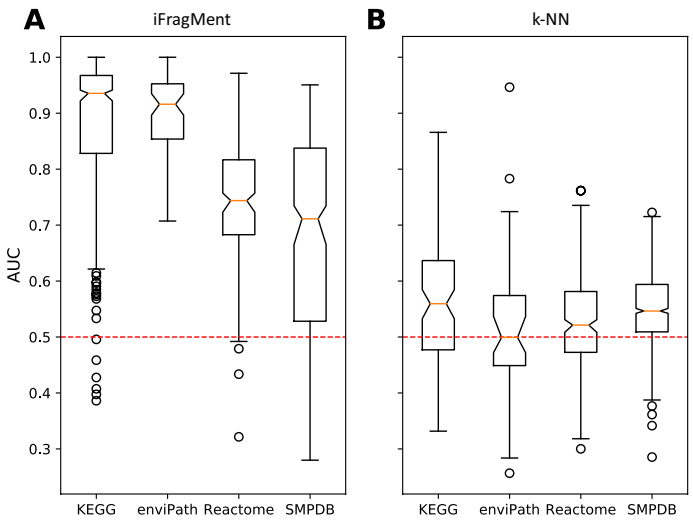

Fig. 1 Distribution of AUC values for the individual pathways in the four databases tested for **a** iFragMent method and **b** k-NN method

Table 2 Global performance (AUC for iFragMent and k-NN) for specific pathways within general KEGG pathway classes

Pathway Class	iFragMent AUC	k-NN AUC	Num. pathways
Drug Development	1.00	0.45	1
Genetic Information Processing	0.94	0.55	2
Metabolism	0.91	0.51	153
Human Diseases	0.82	0.52	10
Environmental Information Processing	0.74	0.52	12
Organismal Systems	0.70	0.50	32
Cellular Processes	0.69	0.47	4

In the following sections, we will analyse several factors that can affect prediction performance: pathway class, compound class (single or multi-pathway) and compound size.

Pathway class

We evaluated the performance of our approach for all KEGG pathways of a given general class (same top-level of the BRITE taxonomy) (Table 2). For that with constructed a ROC curve with the prediction results of all compounds associated with pathways belonging to that class, and calculated the AUC value. Some classes contained a limited number of pathways. Higher AUCs were obtained for 'Drug development' (1.00) (1 pathway: Histamine H2/H3 receptor agonists/antagonists), 'Genetic Information Processing' (0.94) (2 pathways) and 'Metabolism' (0.91) (153 pathways).

Compound size

We also observed that iFragMent performance varies largely depending on the number of chemical fragments present in a compound (a proxy of compound size) (Fig. 2). AUC values of compounds with a small number of fragments (< 32 for KEGG pathways), are

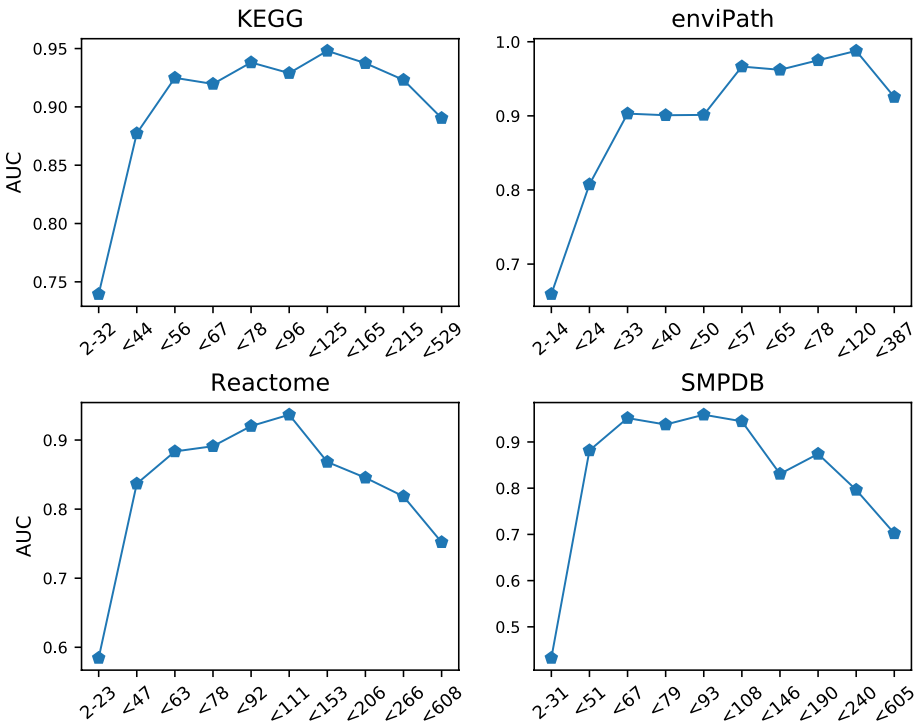

Fig. 2 Performance of iFragMent (AUC) by number of fragments in the four databases tested. Intervals are defined to include a similar number of compounds

Table 3 iFragMent performance (AUC) for compounds involved in a single pathway and those in multiple pathways

Database	AUC uni	AUC multi	Single pathway compounds	Multi pathway compounds	%multi path compounds	AUC global
KEGG	0.95	0.82	4066	1170	23.2	0.88
enviPath	0.94	0.80	1254	91	12.1	0.90
Reactome	0.89	0.70	802	524	43.5	0.74
SMPDB	0.81	0.72	1062	346	35.9	0.74

very low in comparison with that of larger compounds. Similar results are obtained for the other three databases. Performance also decreases (although to a lesser extend) for very large compounds.

Compounds in single versus multiple pathways

We finally assessed the performance of compounds involved in a single pathway and those involved in multiple pathways. Compounds involved in multiple pathways achieved worse results than those involved in a single pathway (Table 3). E.g. for KEGG pathways, single-pathway compounds where predicted with an AUC of 0.95, in contrast to 0.83 AUC for multi-pathway compounds. As the fraction of multi-pathway compounds in each database varies, this partially explained the differences in global AUCs in the four databases studied (Table 3).

Comparison with previous methods

RF-Labute

To compare our approach with that of Machiarulo et al. [8], which uses a Random Forest classifier and Labute descriptors for compounds (RF-Labute), we used their *human_unique* dataset (comprising 52 human pathways) and perform a tenfold cross validaton of our method.

Neither overall performance measures nor numeric results were reported by Machiarulo et al. for their prediction of individual pathways (they did it only for the prediction of pathway classes). Instead, classification errors for individual pathways were graphically shown in a blox-plot organized in seven pathway classes (see Fig. 6D in original publication [8]). Assuming that this figure represents the out-of-bag estimate of error provided by RF, and that this error corresponds to the false discovery rate (FDR), we generated a similar figure for our results (Fig. 3a). As RF-Labute is a multiclass but not a multilabel approach, we evaluated only the top-1 pathway prediction obtained by iFragMent for each compoud.

Visual comparison of both figures (Fig. 3a and Fig. 6D in [8]) reveals higher mean classification errors (FDR) of our method compared to RF-Labute for carbohydrate, lipid, nucleotide and amino acid metabolism pathways (CM, LM, NM and AM); comparable for both energy metabolism and cofactor and vitamins metabolism pathways (EM and CVM); and slightly lower for other amino acids metabolism pathways (OAM).

Part of the errors obtained by iFragMent in the *human_unique* dataset can be due to the small size of some of the pathways, for which reliable fragment statistics could not be obtained (15 out of 52 pathways contained less than 5 compounds, with more than 50% of them with less than 10 compounds).

In addition to the predicted pathway, iFragMent also provides an statistical estimate for each compound-pathway prediction (p value). We have analysed the p values obtained for each of the compound-pathway pairs (as defined in *human_unique*), and

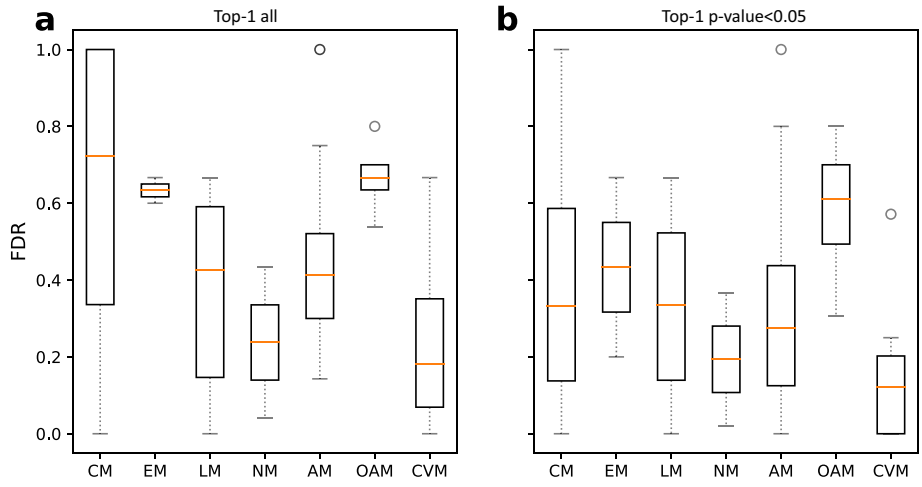

Fig. 3 False discovery rate (FDR) of iFragMent using the *human_unique* dataset from Machiarulo et al. **a** All top-1 predictions. **b** Top-1 predictions with p value < 0.05. CM carbohydrate, EM energy, LM lipid, NM nucleotide, AM amino acid, OAM other amino acids, CVM cofactors and vitamins metabolism

the rank at which they were predicted. As expected, classification errors (FDR) increase as p value increases (Additional file 1: Fig. S1). Thus lower FDR (higher precision or PPV) can be obtained by p value filtering, at the cost of decreasing sensitivy (TPR).

For example, considering all top-1 predictions iFragMent achieved a 0.66 TPR with a classification error (FDR) of 0.34. If we consider only the predictions with p-val < 0.05 in the top-1 positions (Fig. 3b), classification error (FDR) drops to 0.17, at the cost of lowering sensitivity (TPR) to 0.59. Thus, by filtering with p values, we will miss some true compound-pathway associations (mainly those that are not based on enrichment of structural fragments), but make less mistakes.

In the previous section we compared iFragMent top-1 predictions with RF-Labute results. As our approach is a multiclass-multilabel classifier, it allows predicting more than one pathway for a compound. This feature can be exploited to increase the true-positive-rate (TPR, sensitivity). By considering the top-n predictions, TPR increases from 0.66 (top-1) to 0.79 (top-2) and 0.86 (top-3), at the cost of increasing also the number of false positives.

TrackSM

To compare our approach with TrackSM [10], we have designed a real world test scenario where new pathway-compound associations obtained from the release 83.0 of KEGG were predicted. To fairly compare methods, we use the same training dataset provided by TrackSM to construct the iFragMent chemical profiles. Hence, we trained TrackSM and iFragment with the same dataset, and generated predictions for compounds from a more recent version of KEGG. TrackSM reported errors for 41 compounds, that were not included in the comparison. We evaluated a total of 1313 compound-pathway associations involving 127 distinct pathway. Through the analysis of top-1 predictions we obtained a better performance with iFragMent (PPV = 0.41) as compared to TrackSM (PPV = 0.26). (See results in Additional file 1: Table S6).

Chemical substructures associated to a biochemical pathway

Our method allows not only to assign compounds to their biological pathways (hence predicting their biological role) but also to detect the chemical substructures (fragments) that contribute most in such assignments (i.e. the matched components of the pathway Pv vector that are statistically significant).

Figure 4 highlights the three fragments with lowest p values for the KEGG pathway "sphingolipid metabolism" (map00600) on the structures of the compounds associated to this pathway. It can be seen that these three fragments clearly delineate the the hydrocarbon chain and the polar head of sphingosine (C00319), while they are also highlighted in other compounds of the pathway.

Examples

To highlight the advantages as well as the limitations of our approach, we discuss here two examples taken from iFragMent Server, corresponding to predictions of compounds with known biological pathways (see Additional file 1: figures S2-S3).

Correct predictions for histidine (C00135 in KEGG) include 'Histidine metabolism' as the top-ranking pathway, based on the recognition of fragments in the histidine side

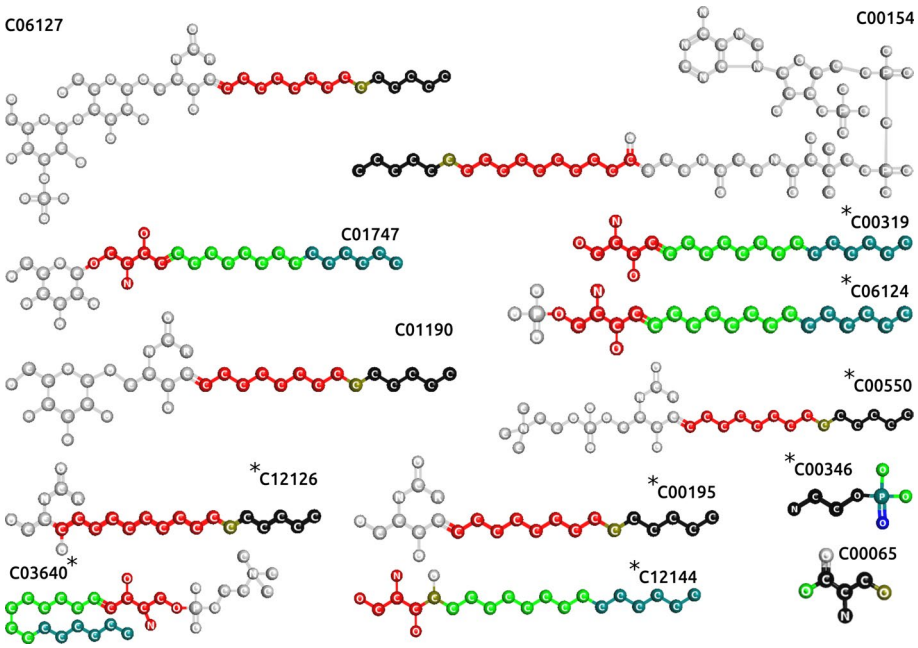

Fig. 4 Enriched fragments (three lowest *p* values) for the "sphingolipid metabolism" compounds (KEGG map00600). From lowest to highest, colored in red, green and blue; their intersection in teal (blue and green), olive (green and blue) and black (red, green and blue)

chain, and 'Biosynthesis of amino acids' based on fragments of the amino and carboxyl functional groups. As these later fragments are common to all amino acids, they also lead to incorrect predictions of other amino acid-related pathways (like 'lysine degradation', 'arginine and proline metabolism' or 'lysine biosynthesis').

iFragMent fails to predict 'Quoring sensing' for phospho-DPD (C20959), the signaling pathway in which it is involved according to the KEGG database. In contrast, it predicts some metabolic and signaling pathways with good *p* values (e.g. 'Glycolysis/Gluconeogenesis', 'Glugacon signaling pathway', 'Pentose phosphate pathway'), based on fragments of its phosphate group attached to the hydrocarbon structure, a substructure commonly found in many metabolic pathways.

Discussion

Profile methods are extensively used in the analysis of biological sequences (proteins, RNA and DNA) [15–18]. They are often used to infer the structure and function of an uncharacterized sequence by its similarity to a group of sequences (used to build the profile), as they detect remote homologies with greater accuracy than pair-wise similarities [19]. Sequence profiles are linear models, built upon the evolutionary relationships established among a group of biological sequences. Although we cannot establish evolutionary relations among chemical structures, chemical profile-inspired approaches have been successfully used in the classification of drug targets [20] relating receptors by ligand similarity, and the prediction of drug 'off-targets' [21] as compounds that bind to a protein have typically similar structures or substructures.

The ability of profile-inspired approaches to assign compounds to their biological pathways has not been explored. As metabolic reactions proceed step-wise, substrate-reactant pairs are structurally related. But given the linear or branched topology of reactions in pathways, we don't know to what extend all compounds of a pathway are structurally related. Indeed, in some cases, structurally similar compounds tend to participate in the same metabolic pathway [22]. Additionally, the recent hypothesis of the "conquest of the chemical space" as the evolutionary driving force of the biological species [23] might suggest that evolution can also be indirectly reflected in chemical networks.

We demonstrate that a profile-inspired approach can be used to predict the potential biological pathway of a chemical compound from its chemical structure alone. We propose a method that relies on enrichment of structural fragments. Although performance varied largely depending on the pathway, we obtained good performances, especially for KEGG metabolic pathways and enviPath biotransformation pathways.

In this work pathways have been defined according to four public databases. KEGG, Reactome and SMPDB include not only metabolic but also other types of pathways such as disease, drug actions, transporting, signaling, etc. Yet, we found some general trends. In all databases, results for compounds involved in a single pathway are much better than for multi-pathway compounds. The percentage of multi-pathway compounds varied largely in the four databases, partially explaining differences in overall database performance. This could be due to the fact that multi-pathway compounds have "mixed" characteristics from more than one pathway, what might confound the predictor. This is equivalent to multi-domain proteins, for example.

Pathways for small compounds (e.g. < 32 fragments in KEGG) were identified with less accuracy than for medium-size compounds. This trend was observed in the four databases analyzed. This could be related to the absence of enough information in their chemical structures to discriminate the pathway they belong to.

The performance of our profile-inspired approach is higher than that obtained for a pair-wise similarity approach (k-NN method) in all databases tested. We also compared our method with the two previously described approaches that addressed the prediction of individual pathways [8, 10]. The three approaches differ both in the structural descriptors used to represent compounds and the algorithms. Our method (iFragMent) and TrackSM are multi-class and multi-label approaches, thus both enable the prediction of more than one pathway for a compound. Labute-RF can only predict one pathway for a compound (i.e. cannot handle multi-pathway compounds).

Testing a reduced set of single-pathway compounds involved in 52 human pathways, Marchiarulo et al. [8] reported lower classification errors than that obtained with iFragMent in its top-1 predictions. Half of the pathways contained less than 10 compounds (bellow the limit set in our work to obtain reliable statistical estimates of fragments). In contrast to machine learning approaches, like Random Forest, we establish a model beforehand to base our predictions: matching pathway-enriched fragments. This allows to provide an statistical estimate (p value) to each prediction, which can be used to decrease classification errors (at the cost of missing some true positives).

In a real-world scenario, where newer compounds never seen by the systems were tested, iFragMent achieved higher PPV than TrakSM [10]. Our method does not consider pathway classes, while TrackSM predicts an individual pathway among those in

a class previously predicted. Errors in pathway class prediction might limit TrackSM performance.

Conclusions

As we accumulate more data on the landscape of small chemicals underlying biological systems, new methodologies are required to mine it so as to extract useful information. Sequence profiles of biological polymers (DNA, RNA and proteins) are behind most approaches that allow mining and interpreting the massive genomic datasets. Consequently, we need similar approaches for the metabolism. A profile-inspired approach grouping functionally related metabolites is useful for assigning new metabolites to the group, as well as for detecting the structural fragments associated to that group (equivalent, for example, to the conserved/functional positions in protein profiles). The last allows getting insight into the chemical basis of the biological activity of a given group of functionally related compounds, in case it is unknown. Both, the assignment of chemical compounds to biological pathways and the detection of the informative fragments using this methodology can be performed by any interested user via the interactive graphical web interface developed.

Methods

Datasets

We compiled pathways and their associated compounds from four resources: KEGG ("pathway" section, release 83.0), Reactome (version v61), SMPDB (release 2.75) and enviPath (EAWAG-BBD dataset, version 0.3.1). Compounds were compiled regardless the organism they are eventually assigned to. We selected those pathways with at least ten chemical compounds, excluding the very general pathways (e.g. from KEGG 'Metabolic pathways' (map01100); 'Biosynthesis of secondary metabolites' (map01110); 'Microbial metabolism in diverse environments' (map01120); 'Biosynthesis of antibiotics' (map01130) and 'Degradation of aromatic compounds' (map01220)).

Structural data files of compounds were downloaded from their respective database for KEGG, EnviPath and SMPDB pathways. For Reactome pathways, we retrieved chemical structure files from ChEBI [24] using the cross-references provided in Reactome. Datasets (excepting KEGG, due to license requirements) are available at https://github.com/jlopez-ibanez/iFragMent.

Compound structural descriptors

From their structural data file, we generated a vectorial representation of compounds using the molecular fragments obtained with ISIDA Fragmentor [25]. We generated all linear fragments of 1–7 atoms and all atom centered fragments of 2–4 atoms (ISIDA Fragmentor parameters -t3 -l2 -u7 -t6 -u4 -t0). We then coded each chemical compound as a binary vector, where each vector component represents the presence or absence of a given fragment. The vector length depends on the database analyzed (Additional file 1: Table S1).

In some cases, we found compounds with the same vectorial representation. These redundant compounds were merged and treated as a single compound in the evaluation

of the method. This avoids identical compounds to be present both in the training and test sets.

Scoring

From the vectorial representation of the compounds in a given pathway, we obtain a vector representation for that pathway (Pv) (Fig. 5). Each vector component corresponds to a fragment in the database, and represents the probability (Pv) of observing it by chance in the compounds of the pathway, compared to a background distribution (all the compounds in the database). For each fragment and pathway, p values are calculated with the cumulative hypergeometric distribution:

$$Pv = \sum_{i=x}^{N} \frac{\binom{K}{i}\binom{M-K}{N-i}}{\binom{M}{N}} \tag{1}$$

where M is the total number of compounds in the database, K is the total number of compounds in the database with the fragment, N is the number of compounds in the pathway and x is the number of compounds in the pathway with the fragment.

With this pathway representation, we define a score for quantifying the matching of a given compound (X) against a pathway P, Sc, as:

$$Sc = \frac{-\sum_{i=1}^{n} \log\left(Pv_i\right) \cdot X_i}{\sum_{i=1}^{n} X_i} \tag{2}$$

where Xi is the ith component in the fingerprint representation of compound X, and Pv is the probability vector of the presence of fragments in the pathway as defined in Eq. 1.

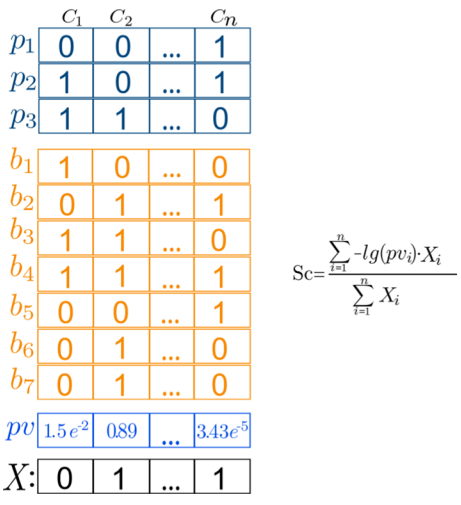

Fig. 5 Example of calculating the score of a query compound (X) against a profile using iFragMent. Columns (f_i) represent fragments, rows represent compounds: in blue (p_i) those involved in pathway (P) and orange (b_i) those that are not (background). Presence/absence is coded with 1/0. Vector pv contains the probability of observing a fragment in the compounds of the pathway by chance, considering their distribution in the background. The score of X against P (Sc) is calculated as shown

To avoid log(0), those fragments with a p value equal to zero were assigned a probability equal to the smallest p value obtained divided by 50. Note that only fragments present in compound x (i.e. with a "1" in the vectorial representation) are taken into account with this formulation.

We observed, as expected, that scores Sc of compounds belonging to a pathway P tend to be higher than the scores of compounds not associated with that pathway (even if we exclude the compound to calculate Pv) (data not shown). However, scores of the same compound in different pathways are not comparable, so they can not be directly used as to predict in which pathway a compound participates. To solve this, we devised a random statistical model.

Statistical model

In order to compare and rank the scores of a compound X in different pathways we calculate the corresponding z-scores using a null model, using a similar approach to [26]. For each pathway, we obtained a random distribution of scores (Sr). For that, we performed 100,000 randomizations of the matrix (NxM) of N compounds and M fingerprints (being N all the compounds in the corresponding database), and scored the resulting N randomized compound fingerprints against the pathway. Mean and standard deviations of Sr were obtained, and used to calculate z-scores from scores (Sc). We checked that Sr followed a extreme value distribution, as in [26]. Parameters characterizing each random distribution of scores (scale and location) were calculated with evfit function (MATLAB 2010b) and used to analytically calculate p values from z-scores.

To predict the pathway of a compound, we calculate the z-scores and corresponding p values of the compound against all pathway pv models. We then rank the pathway p values by increasing value. We take those pathways in top-n ranking positions as the predictions, creating a whole family of predictors with increasing sensitivity and decreasing specificity. A Receiving Operating Characteristic curve (ROC curve) is calculated as n increases from 1 to the total number of pathways. Overall performance was quantified as the area under the ROC curve (AUC). A value of AUC = 0.5 would represent a random prediction (correct and incorrect pathways uniformly distributed in the ranked list), while values higher than that represent good predictions (correct pathways closer to the top of the list).

Both p values and z-scores are reported by the iFragMent web server. The chemical substructures that are characteristic (enriched) in a biological pathway are obtained from the pathway Pv vector as those with the highest enrichment in that pathway (i.e. lowest p value). These are used to highlight matched enriched fragments in the query compounds.

K-nearest neighbour method

We implemented a k-nearest neighbour (k-NN) approach using the same vector representation of compounds. We calculate the structural similarity of a query compound to all the compounds in the database using the Tanimoto coefficient. We finally assign the pathways of the top-k most similar compounds. ROC curves and AUC values are calculated as previously explained.

Evaluation

We evaluated our method, as well as the k-NN approach, with a tenfold cross validation, using exactly the same partition of training datasets for both methods.

Comparison with previous approaches

Dataset *human_unique* was downloaded from the Supplementary material of Marchiarulo et al. [8]. We used this dataset to perform a tenfold cross validation of our method, and compare results with those reported in [8]. TrackSM software [10] was downloaded from https://dna.engr.uconn.edu/?page_id=648. TrackSM training dataset was obtained from the Config directory files, and were used to calculate iFragMent profiles, so that both methods can be compared trained in the same sets. Novel compounds in KEGG (release 83.0) not included in this TrackSM training dataset were used to predict pathways with both TraskSM and iFragMent.

Abbreviations

AUC: Area under the ROC curve; FDR: False discovery rate; k-NN: K-nearest neighbour; PPV: Positive predictive value or precision; RF: Random forest; ROC: Receiving operating characteristic curve; TPR: True positive rate.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04252-y.

Additional file 1: Table S1. Composition of databases and fingerprints used. **Table S6**. Precision (or positive predictive value, PPV) and sensitivity of the prediction of a dataset of 1313 compounds using TrackSM and iFragMent methods. **Figure S1**. The associated p-value resulting from iFragMent predictions can be used to reduce the FDR. **Figure S2**. iFragMent server predictions for histidine (C00135 in KEGG). **Figure S3**. iFragMent server predictions for phospho-DPD (C20959 in KEGG).

Additional file 2: Table S2. AUC results for individual pathways in KEGG dataset. **Table S3**. AUC results for individual pathways in enviPath dataset. **Table S4**. AUC results for individual pathways in Reactome dataset. **Table S5**. AUC results for individual pathways in SMPDB dataset.

Acknowledgements

Not applicable.

Authors' contributions

JLI performed the analysis and developed the code and web server. MC conceived the work and drafted the manuscript. JLI, FP and MC developed the method, analyzed results and wrote the final version of the manuscript. All authors have read and approved the final manuscript.

Funding

This work was partially funded by the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) (project BIO2015-72091-EXP). J.L.I was the recipient of a pre-doctoral contract from the MINECO and European Social Fund (BIO2010-22109). The funding agencies played no additional role in the research and preparation of this manuscript.

Availability of data and materials

Code and datasets analysed during the current study (except KEGG data) are available in the GitHub repository, https://github.com/jlopez-ibanez/iFragMent. The KEGG datasets analysed were accessed under a license that restricted their re-distrubution. KEGG data access is available from Kanehisa Laboratories https://www.kegg.jp/kegg/download/.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Florencio Pazos is member of the editorial board (Associate Editor).

Received: 15 April 2021 Accepted: 9 June 2021
Published online: 12 June 2021

References

1. Dobson CM. Chemical space and biology. Nature. 2004;432(7019):824–8. https://doi.org/10.1038/nature03192.
2. Cai YD, Qian Z, Lu L, Feng KY, Meng X, Niu B, et al. Prediction of compounds' biological function (metabolic pathways) based on functional group composition. Mol Divers. 2008;12(2):131–7. https://doi.org/10.1007/s11030-008-9085-9.
3. Lu J, Niu B, Liu L, Lu WC, Cai YD. Prediction of small molecules' metabolic pathways based on functional group composition. Protein Pept Lett. 2009;16(8):969–76.
4. Hu LL, Chen C, Huang T, Cai YD, Chou KC. Predicting biological functions of compounds based on chemical-chemical interactions. PLoS ONE. 2011;6(12): e29491. https://doi.org/10.1371/journal.pone.0029491.
5. Gao YF, Chen L, Cai YD, Feng KY, Huang T, Jiang Y. Predicting metabolic pathways of small molecules and enzymes based on interaction information of chemicals and proteins. PLoS ONE. 2012;7(9): e45944. https://doi.org/10.1371/journal.pone.0045944.
6. Chen L, Chu C, Feng K. Predicting the types of metabolic pathway of compounds using molecular fragments and sequential minimal optimization. Comb Chem High Throughput Screen. 2016;19(2):136–43.
7. Baranwal M, Magner A, Elvati P, Saldinger J, Violi A, Hero AO. A deep learning architecture for metabolic pathway prediction. Bioinformatics. 2020;36(8):2547–53.
8. Macchiarulo A, Thornton JM, Nobeli I. Mapping human metabolic pathways in the small molecule chemical space. J Chem Inf Model. 2009;49(10):2272–89. https://doi.org/10.1021/ci900196u.
9. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353–61. https://doi.org/10.1093/nar/gkw1092.
10. Hamdalla MA, Rajasekaran S, Grant DF, Mandoiu II. Metabolic pathway predictions for metabolomics: a molecular structure matching approach. J Chem Inf Model. 2015;55(3):709–18. https://doi.org/10.1021/ci500517v.
11. Durbin R, Eddy SR, Krogh A, Mitchison G. Biological sequence analysis. Cambridge: Cambridge University Press; 1998.
12. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2017. https://doi.org/10.1093/nar/gkx1132.
13. Jewison T, Su Y, Disfany FM, Liang Y, Knox C, Maciejewski A, et al. SMPDB 2.0: big improvements to the small molecule pathway database. Nucleic Acids Res. 2014;42(Database issue):D478–84. https://doi.org/10.1093/nar/gkt1067.
14. Wicker J, Lorsbach T, Gutlein M, Schmid E, Latino D, Kramer S, et al. enviPath–The environmental contaminant biotransformation pathway resource. Nucleic Acids Res. 2016;44(D1):D502–8. https://doi.org/10.1093/nar/gkv1229.
15. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A. Use of the "Perceptron" algorithm to distinguish translational initiation sites in *E. coli*. Nucleic Acids Res. 1982;10(9):2997–3011.
16. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA. 1987;84(13):4355–8.
17. Brown M, Hughey R, Krogh A, Mian IS, Sjolander K, Haussler D. Using Dirichlet mixture priors to derive hidden Markov models for protein families. Proc Int Conf Intell Syst Mol Biol. 1993;1:47–55.
18. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol. 1994;235(5):1501–31. https://doi.org/10.1006/jmbi.1994.1104.
19. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, et al. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol. 1998;284(4):1201–10. https://doi.org/10.1006/jmbi.1998.2221.
20. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, et al. Predicting new molecular targets for known drugs. Nature. 2009;462(7270):175–81. https://doi.org/10.1038/nature08506.
21. Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, et al. Large-scale prediction and testing of drug activity on side-effect targets. Nature. 2012;486(7403):361–7. https://doi.org/10.1038/nature11159.
22. Hattori M, Okuno Y, Goto S, Kanehisa M. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. J Am Chem Soc. 2003;125(39):11853–65. https://doi.org/10.1021/ja036030u.
23. de Lorenzo V. From the selfish gene to selfish metabolism: revisiting the central dogma. BioEssays. 2014;36(3):226–35. https://doi.org/10.1002/bies.201300153.
24. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, et al. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res. 2016;44(D1):D1214–9. https://doi.org/10.1093/nar/gkv1031.
25. Ruggiu F, Marcou G, Varnek A, Horvath D. ISIDA Property-labelled fragment descriptors. Mol Inform. 2010;29(12):855–68. https://doi.org/10.1002/minf.201000099.
26. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. Nat Biotechnol. 2007;25(2):197–206.

Publisher's Note