J Math Chem (2009) 45:598–606
DOI 10.1007/s10910-008-9348-9

ORIGINAL PAPER

# Parameter-free linear relationship (PFLR) and its application to 3D QSAR

Ödön Farkas · Imre Jákli · Adrián Kalászi · Gábor Imre

Received: 14 November 2007 / Accepted: 27 November 2007 / Published online: 24 January 2008

**Abstract** The linear relationship is still the most important tool for establishing connection between correlating features, properties. The name "parameter-free linear relationship" (PFLR) stands for a new formalism, a generalized interpolation scheme, which can be readily used for predicting biological activities or other properties in 3D QSAR manner. Our studies demonstrate the good predictive power of PFLR even when used with a simple set of 3D molecular descriptors without constructing a grid representation of the features. PFLR allows completing most of the computations solely in the space of descriptors, without experimental training data, which, however, bears no importance in the case of 3D QSAR but might be advantageous in other areas where multidimensional linear regression or partial least squares based methods are applicable.

**Keywords** PFLR · Parameter-free · Linear relationship · 3D QSAR · QSPR · Linear regression · Multidimensional interpolation · Partial least squares · PLS

## 1 Introduction

There is an abundance of applications of linear relationship, implemented in methods based on interpolation, linear regression, (partial) least squares (PLS) [1]. Curiously, one of the very successful interpolation schemes, the direct inversion in the iterative subspace or DIIS method [2,3] of Pulay, widely used in quantum chemistry to accelerate SCF convergence [2,3] but practically unknown in other areas except it's geometry optimization version the GDIIS [4,5]. The reason of such ignorance could

Ö. Farkas (✉) · I. Jákli · A. Kalászi · G. Imre
Laboratory of Chemical Informatics, Institute of Chemistry, Eötvös Loránd University,
1/A Pázmány Péter sétány, 1117 Budapest, Hungary
e-mail: farkas@chem.elte.hu

be due to the specialized area and also because the interpretation wrapped the more generic potential of the method. DIIS and GDIIS are even regarded erroneously as "heuristic" methods. The same idea is also appears in the Convex Constraint Analysis, CCA, algorithm, Perczel et al. [6–9] for decomposing spectra. The proposed method, the parameter-free linear relationship or PFLR can be regarded as a generalization of Pulay's DIIS method.

The proposed method can be a viable alternative of any other approach based on assuming a linear relationship but has been developed for the purpose of building up an automatic prediction scheme using 3D descriptors for QSAR. Commercially available, well established methods, like CoMFA [10] and CoMSIA [11,12] are furnished with graphical user interfaces to facilitate the otherwise not too simple process of building up a 3D QSAR model. The automatic process also has to deal with many tasks, like automatic generation and selection of the potential bioactive conformation, optimal choice of training molecules, superimposing molecules, adjusting model parameters, etc. The pieces of such process should be validated one-by-one, the current study is focused on introducing and validating the prediction engine. The introductory papers of different methods usually contain sets with a few external test molecules and are not suitable for thorough comparison. Also, as the well established methods are only commercially accessible, we used the data published by Sutherland et al. [13] in their detailed comparison of the predictive power of different QSAR methods and models. The importance of the proper (automatic) alignment will also be demonstrated.

## 2 Method

### 2.1 PFLR formulae

We need only one presumption for PFLR to work, namely, the changes of independent variables are proportional to the changes of dependent variables:

$$\boldsymbol{\Delta}\mathbf{x} \sim \Delta y \tag{1}$$

As it is not easy to introduce fully generic notations for independent and dependent variables, therefore, for the sake of simplicity, independent variables are regarded to be vectors, $\mathbf{x}$, and the dependent variables are regarded to be scalars, $y$, like a set of descriptors and the corresponding activities in a QSAR study. However, it is important to note here that such restrictions are not necessary for the method to work. Also, due to the symmetry of the linear relationship, the choice of "dependent" and "independent" side is arbitrary. If a number of independent variable values and the corresponding dependent variables are known then the same linear combination of observed changes of independent variables and observed changes of dependent variables will result corresponding changes at both side:

$$\sum_{i=1}^{n} \kappa_i \left(\mathbf{x}_i - \mathbf{x}_0\right) = \Delta\mathbf{x} \Leftrightarrow \sum_{i=1}^{n} \kappa_i \left(y_i - y_0\right) = \Delta y \tag{2}$$

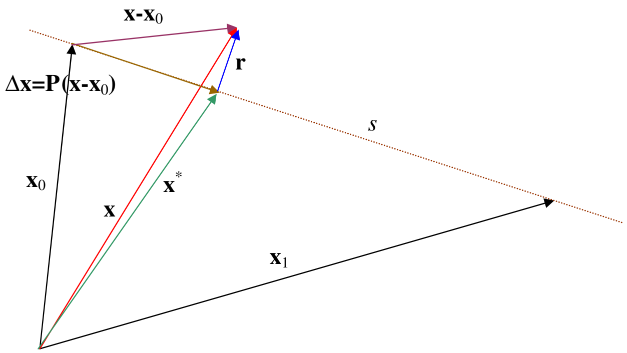

**Fig. 1** The figure illustrates in two dimensions how to obtain the $\mathbf{x}^*$ approximation of the arbitrary independent variable vector, $\mathbf{x}$ . The change vector, $\mathbf{x} - \mathbf{x}_0$ (pointing from the representative point, $\mathbf{x}_0$, to the arbitrary point, $\mathbf{x}$), is projected into the $s$ subspace of available change vectors, $\mathbf{x}_i - \mathbf{x}_0$, using a projector, $\mathbf{P}$. The error of the approximation, the *residuum*, $\mathbf{r}$, is perpendicular to the subspace of change vectors

where the $\Delta y$ change of dependent variables corresponds to the $\mathbf{\Delta x}$ change of independent variables. The changes in Eq. (2) are measured from chosen representative points, $\mathbf{x}_0$ and $y_0$, which is also not necessary for Eq. (2) to hold but used in the following steps. The main idea is to approximate the independent variables using the subspace of available variable changes and use the obtained coefficients to express the corresponding dependent variable.

Therefore, goal of PFLR is to approximate an arbitrary independent variable, $\mathbf{x}$, using a representative point, $\mathbf{x}_0$, and a linear combination of the available variable changes, $\mathbf{x}_i - \mathbf{x}_0$. For that purpose, the difference vector pointing from the representative point to the arbitrary point, $\mathbf{x} - \mathbf{x}_0$, is projected into the subspace of available independent variable changes and the projection is expressed in terms of a linear combination of variable changes (see Fig. 1). The corresponding projector can be expressed in terms of the variable change vectors:

$$\mathbf{P} = \mathbf{X}_{\mathbf{\Delta}} \mathbf{X}_{\mathbf{\Delta}}^{-} \tag{3}$$

where $\mathbf{X}_{\mathbf{\Delta}}$ denotes the matrix collecting the available independent variable changes, $\mathbf{x}_i - \mathbf{x}_0$, in its columns, while "–" stands for generalized inverse which can be constructed *via* singular value decomposition, SVD. The expression of the whole approximation follows as:

$$\begin{aligned}
\mathbf{x}^* &= \mathbf{x}_0 + \mathbf{P}\left(\mathbf{x} - \mathbf{x}_0\right) = \mathbf{x}_0 + \mathbf{X}_{\mathbf{\Delta}} \mathbf{X}_{\mathbf{\Delta}}^{-}\left(\mathbf{x} - \mathbf{x}_0\right) \\
\mathbf{k} &= \mathbf{X}_{\mathbf{\Delta}}^{-}\left(\mathbf{x} - \mathbf{x}_0\right) \\
\mathbf{x}^* &= \mathbf{x}_0 + \mathbf{X}_{\mathbf{\Delta}} \mathbf{k} = \mathbf{x}_0 + \sum_{i=1}^{n} \kappa_i \left(\mathbf{x}_i - \mathbf{x}_0\right)
\end{aligned} \tag{4}$$

where vector $\mathbf{k}$ collects the $\kappa_i$ coefficients of the variable changes in demand.

The representative point, $\mathbf{x}_0$, is one of the observed independent variable vectors or a special linear combination of them. As it will be demonstrated below, if the representative point is the average of the available independent variable vectors than PFLR provides identical results to linear regression, therefore, this choice was used throughout the applications. The choice of

$$\mathbf{x}_0 = \sum_{i=1}^{n} \frac{1}{n} \mathbf{x}_i \tag{5}$$

leads to the following $c_i$ coefficients on the independent variable vectors:

$$\mathbf{x}_0 + \sum_{i=1}^{n} \kappa_i \left( \mathbf{x}_i - \mathbf{x}_0 \right) = \sum_{i=1}^{n} \left[ \left( \frac{1 - \sum\limits_{i=1}^{n} \kappa_i}{n} + \kappa_i \right) \mathbf{x}_i \right]$$

$$c_i = \frac{1 - \sum\limits_{i=1}^{n} \kappa_i}{n} + \kappa_i \Rightarrow \sum_{i=1}^{n} c_i = 1 \tag{6}$$

The $\sum\limits_{i=1}^{n} c_i = 1$ condition ensures on its own that the corresponding linear combination will result in a vector in the subspace spanned by the endpoints of the independent variable vectors. The representative point, $\mathbf{x}_0$, in general, should be a point of the same subspace.

### 2.2 PFLR and linear regression

The connection between PFLR and linear regression is demonstrated here *via* a simple one–one dimensional linear fit as it shown on Fig. 2.

### 2.3 3D QSAR descriptors

The independent variables in case of QSAR studies are the molecular descriptors or scores. One well known solution is to describe molecular properties as grid points around the molecule. Such approach was introduced for the CoMFA, CoMSIA family of methods. As the interpolation does not need an explicitly formed model, we have decided to represent the features at full accuracy, using continuous functions. Continuous functions can form a Hilbert space, an abstract vector space, using the following definition for the inner product

$$\langle a, b \rangle = \iiint\limits_{V} \phi_a \phi_b \tag{7}$$

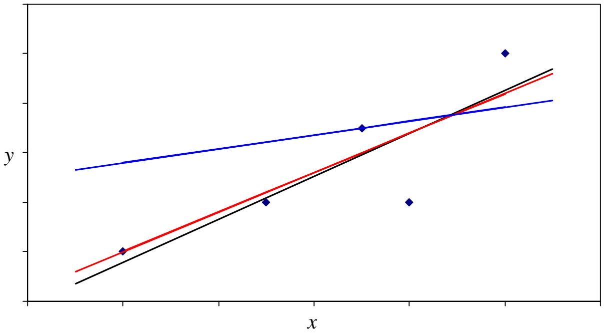

**Fig. 2** The figure demonstrates possible choices for the representative point, $\mathbf{x}_0$. The lines are defined by the independent variable, $x$, and the corresponding prediction, $y$. When the representative point is one of the given independent variable values then the prediction line goes through the corresponding $(x_i, y_i)$ point. If the representative point is the average of the given points then the prediction gives identical results to unconstrained line fit

which is necessary and enough to compute the $c_i$ PFLR coefficients for the purpose of predicting activities or other properties.

The descriptors, $\phi_a$ and $\phi_b$ are functions and the inner product should be calculated after proper alignment. The features are represented as the sum of atom-centered Gaussian functions (see Fig. 3). For each atom, if the actual feature is present then the sum of the four Gaussian functions is added to the molecular descriptor function. The descriptors are the existence of an atom (steric), six regular pharmacophore types, as provided by the fragment based algorithm of ChemAxon's JChem package [14]. In case of charges, the atomic contribution is scaled by the atomic charge as assigned by ChemAxon Marvin's [15] charge plugin. The contribution of certain features are also scaled according to the model's setup. The final prediction is composed as the average of the predicted values provided by the three predefined models (see Table 1). The predefined models consist of two simple, steric and charge, and one composite, using all pharmacophore types and steric contribution. The aromatic pharmacophore type is

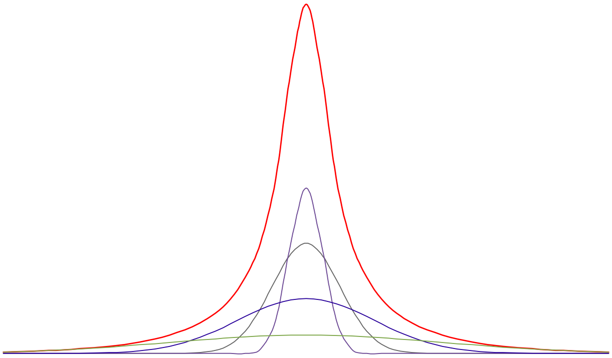

**Fig. 3** If a feature is present at an atom than it is represented as the sum of four atom-centered 3D Gaussian functions. The sum of such feature-representing functions is used as molecular descriptor

**Table 1** The weights of different features in the descriptors used by the 3D QSAR models

| Features | Model #1 | Model #2 | Model #3 |
|---|---|---|---|
| Steric | 1.0 | – | 2.0 |
| Charge | – | 1.0 | – |
| + | – | – | 1.0 |
| — | – | – | 1.0 |
| H-bond donor | – | – | 1.0 |
| H-bond acceptor | – | – | 1.0 |
| Hydrophobic | – | – | 1.0 |
| Aromatic | – | – | 0.25 |

usually assigned to complete rings, therefore, a weight of 0.25 was used for the atomic contributions. As the aromatic rings are represented by their atoms no extra care was taken to describe their orientation. Commercial models usually contain counter parts for the charged atoms and the H-bond acceptors and donors. We assigned the corresponding pharmacophore types for the neighbor atoms to describe the orientation of these features. The protonation state of the molecules may change while binding to an active site, therefore the union of features for all microspecies was assigned to the descriptors. The microspecies were generated *via* the $\mathrm{pK_a}$ plugin of ChemAxon's Marvin [15].

The best stand-alone prediction was observed with the composite model but still, the average result of all three models provided the best *overall* prediction quality and was selected for further use. As the model building process needs only the computation of the overlap (inner product) of the descriptors, an individual selection of the most similar training molecules for the purpose of more accurate prediction is feasible. This, "Local PFLR", L-PFLR, version of the method chooses the training molecules which have closer descriptors to the actual subject of prediction than the standard deviation of all training molecules measured from it.

## 3 Results and discussion

Comparisons of 3D QSAR methods often use results with different sources of 3D molecular structures, different selection of conformations or different method for alignment while each of these differences can strongly affect the prediction strength of 3D QSAR model. Also, a small number of test molecules may not be satisfactory for statistical analysis. The detailed and thorough comparison of Sutherland et al. [13] provides all necessary data for strict comparison of the predictive strength of 3D QSAR methods. The source of data and the alignment procedures are described in their paper's supplementary material. Important to note, that they used the average of training set activities for computing predictive $r^2$ values and we also published our results accordingly, for the sake of correct comparison, otherwise, our implementation normally provides the less favorable predictive $r^2$ using the average of the test activities. The comparison can be found in Table 2. The bold values indicate the best results and the results within 10% compared to the best one. In short, CoMSIA/extra, PFLR and L-PFLR gave considerably better results than CoMFA or CoMSIA/basic

**Table 2** The table summarizes the PFLR-QSAR prediction results compared to CoMFA [10] and CoMSIA [11,12]

| Set[a] | Number of molecules | | CoMFA | | CoMSIA basic | | CoMSIA extra | | PFLR | | L-PFLR | | L-PFLR + alignment[c] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | training | test | $r^2$ | $s$ | $r^2$ | $s$ | $r^2$ | $s$ | $r^2$ | $s$ | $r^2$ | $s$ | $r^2$ | $s$ |
| ACE | 76 | 38 | 0.49 | 1.54 | **0.52** | **1.48** | 0.49 | **1.53** | **0.53** | **1.45** | **0.57** | **1.40** | 0.68 | 1.21 |
| AchE | 74 | 37 | 0.47 | 0.95 | 0.44 | 0.98 | 0.44 | 0.98 | **0.57** | **0.85** | **0.62** | **0.80** | 0.58 | 0.84 |
| BZR | 98 | 49 | 0.00 | 0.97 | 0.08 | **0.93** | 0.12 | **0.91** | 0.15 | **0.89** | **0.22** | **0.85** | 0.25 | 0.83 |
| COX2 | 188 | 94 | 0.29 | **1.24** | 0.03 | 1.44 | **0.37** | **1.17** | 0.26 | **1.26** | 0.28 | **1.24** | 0.28 | 1.24 |
| DHFR | 237 | 124 | **0.59** | **0.89** | 0.52 | **0.96** | **0.53** | **0.95** | **0.55** | **0.93** | **0.58** | **0.89** | 0.65 | 0.81 |
| GPB | 44 | 22 | 0.42 | 0.94 | 0.46 | 0.90 | **0.59** | **0.79** | 0.52 | **0.83** | **0.58** | **0.78** | 0.56 | 0.80 |
| THER | 51 | 25 | **0.54** | **1.59** | 0.36 | 1.87 | **0.53** | **1.60** | 0.41 | **1.75** | **0.49** | **1.63** | 0.26 | 1.97 |
| THR | 59 | 29 | **0.63** | **0.70** | 0.55 | **0.76** | **0.63** | **0.70** | **0.63** | **0.69** | **0.64** | **0.69** | 0.63 | 0.68 |
| Overall/weighted[b] | 827 | 418 | 0.42 | 1.07 | 0.34 | 1.15 | 0.44 | 1.06 | 0.43 | 1.06 | 0.47 | 1.03 | 0.49 | 1.01 |

Then name of the test set, number of training and test molecules, $r^2$ and standard deviation ($s$) values for the compared methods are shown. Bold values indicate a result which is in 10% vicinity of the best achieved value. Local PFLR was also tested using our own alignment method ("Local PFLR + alignment") to demonstrate the importance of the alignment of molecules for 3D QSAR studies

[a] The aligned test molecules, experimental activities, CoMFA and CoMSIA results were taken from ref. [13]

[b] Total number of training/test molecules and weighted average $r^2$ and s values. The data is weighted by the number of test molecules

[c] Local PFLR prediction results using our built-in alignment module. These data were excluded when the best results were chosen

and L-PFLR gave slightly better results than all others. It is important to note, that except the choice of the local or full version of PFLR, there was no user adjustable parameter used for the model building and prediction process of our method. The importance of alignment is demonstrated in the *last* column, where our automatic alignment procedure was also utilized. In one case, the THERM set, our alignment method almost completely failed but usually helped to reach better results than the original alignment.

## 4 Summary

The PFLR interpolation scheme is a viable alternative of multivariate linear regression or PLS, specially, in cases, when the fast process of repeated predictions using the same descriptors, or independent variables is required. It was demonstrated that the PFLR interpolation can give identical results to linear regression; however, the complete proof will be given later.

We constructed a simple representation of molecular features, like pharmacophore type, atomic charge and steric hindrance, to facilitate the application of PFLR to 3D QSAR. The success of the method was demonstrated *via* a comparison to well established, commercially available methods, CoMFA and CoMSIA using data found in the literature [13]. The method will be available through ChemAxon (http://www.chemaxon.com) as part of their Instant JChem product [16].

**Acknowledgements** This work was supported by the Chemaxon Ltd., the Infopark Foundation, and the Hungarian Applied Research Grant, GVOP-3.1.1.-2004-05-0451/3.0. The authors thank the fruitful discussions to Ildikó Magdó of Richter Gedeon Ltd., Ferenc Csizmadia, Ákos Papp and Miklós Vargyas of ChemAxon Ltd.

## References

1. S. Wold, M. Sjostrom, L. Eriksson, Pls-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Sys. **58**(2), 109–130 (2001)
2. P. Pulay, Convergence acceleration of iterative sequences—the case of scf iteration. Chem. Phys. Lett. **73**(2), 393–398 (1980)
3. P. Pulay, Improved scf convergence acceleration. J. Comput. Chem. **3**(4), 556–560 (1982)
4. P. Csaszar, P. Pulay, Geometry optimization by direct inversion in the iterative subspace. J. Mol. Struct. **114**, 31–34 (1984)
5. O. Farkas, H.B. Schlegel, Methods for optimizing large molecules - part iii. An improved algorithm for geometry optimization using direct inversion in the iterative subspace (GDIIS). Phys. Chem. Phys. **4**, 11–15 (2002)
6. A. Perczel, M. Hollosi, G. Tusnady, G.D. Fasman, Convex constraint decomposition of circular-dichroism curves of proteins. Croatica. Chemica. Acta **62**(2A), 189–200 (1989)
7. A. Perczel, M. Hollosi, G. Tusnady, G.D. Fasman, Convex constraint analysis—a natural deconvolution of circular-dichroism curves of proteins. Protein Eng. **4**(6), 669–679 (1991)
8. A. Perczel, K. Park, G.D. Fasman, Deconvolution of the circular-dichroism spectra of proteins—the circular-dichroism spectra of the antiparallel beta-sheet in proteins. Proteins **13**(1), 57–69 (1992)
9. A. Perczel, K. Park, G.D. Fasman, Analysis of the circular-dichroism spectrum of proteins using the convex constraint algorithm—a practical guide. Anal. Biochem. **203**(1), 83–93 (1992)
10. R.D. Cramer, D.E. Patterson, J.D. Bunce, Comparative molecular-field analysis (comfa). 1. Effect of shape on binding of steroids to carrier proteins. J. Am. Chem. Soc. **110**(18), 5959–5967 (1988)

11. G. Klebe, U. Abraham, T. Mietzner, Molecular similarity indexes in a comparative-analysis (comsia) of drug molecules to correlate and predict their biological-activity. J. Med. Chem. **37**(24), 4130–4146 (1994)
12. G. Klebe, U. Abraham, Comparative molecular similarity index analysis (comsia) to study hydrogen-bonding properties and to score combinatorial libraries. J. Comput.-Aided Mol. Des. **13**(1), 1–10 (1999)
13. J.J. Sutherland, L.A. O'Brien, D.F. Weaver, A comparison of methods for modeling quantitative structure-activity relationships. J Med. Chem. **47**(22), 5541–5554 (2004)
14. ChemAxon, "JChem Base," F. Csizmadia (Editor) (2007), http://www.chemaxon.com/product/jc_base.html, Budapest
15. ChemAxon, "Marvin," F. Csizmadia (Editor) (2007), http://www.chemaxon.com/product/marvin_land.html, Budapest
16. ChemAxon, "Instant JChem," F. Csizmadia (Editor) (2007), http://www.chemaxon.com/product/ijc.html, Budapest