Fe(2)OG: an integrated HMM profile-based web server to predict and analyze putative non-haem iron(II)- and 2-oxoglutarate-dependent dioxygenase function in protein sequences

Siddhartha Kundu*

Abstract
Objective: Non-haem iron(II)- and 2-oxoglutarate-dependent dioxygenases (i2OGdd), are a taxonomically and functionally diverse group of enzymes. The active site comprises ferrous iron in a hexa-coordinated distorted octahedron with the apoenzyme, 2-oxoglutarate and a displaceable water molecule. Current information on novel i2OGdd members is sparse and relies on computationally-derived annotation schema. The dissimilar amino acid composition and variable active site geometry thereof, results in differing reaction chemistries amongst i2OGdd members. An additional need of researchers is a curated list of sequences with putative i2OGdd function which can be probed further for empirical data.

Results: This work reports the implementation of Fe(2)OG, a web server with dual functionality and an extension of previous work on i2OGdd enzymes (Fe(2)OG ≡ {H2OGpred, DB2OG}). Fe(2)OG, in this form is completely revised, updated (URL, scripts, repository) and will strengthen the knowledge base of investigators on i2OGdd biochemistry and function. Fe(2)OG, utilizes the superior predictive propensity of HMM-profiles of laboratory validated i2OGdd members to predict probable active site geometries in user-defined protein sequences. Fe(2)OG, also provides researchers with a pre-compiled list of analyzed and searchable i2OGdd-like sequences, many of which may be clinically relevant. Fe(2)OG, is freely available (http://204.152.217.16/Fe2OG.html) and supersedes all previous versions, i.e., H2OGpred, DB2OG.

Keywords: Facial triad, Hidden markov model, Non-haem iron(II)- and 2-oxoglutarate-dependent dioxygenases

Introduction
Dioxygenases, unlike monooxygenases are oxidoreductases which can incorporate both atoms of molecular oxygen, one each into a substrate and co-substrate (Fig. 1). These enzymes are classified on the basis of a metal co-factor (iron, cobalt, nickel, copper) and the presence of a haem-prosthetic group. Iron-based dioxygenases are mononuclear (extradiol catechol, EC1.13.11,x; 2-oxoglutarate-dependent, EC1.14.11,x), possess Rieske clusters (naphthalene 1,2-dioxygenase, EC1.14.12,12) and may utilize haem (indoleamine 2,3-dioxygenase, EC1.13.11,52; tryptophan 2,3-dioxygenase, EC1.13.11,11) [1–5]. Extradiol and 2OG-dependent dioxygenases, possess a triad of catalytically competent $HX_n[DE]X_nH$ residues and comprise one face of a distorted octahedral co-ordination...
sphere with iron(II) [3–6]. The other face is formed by three displaceable water molecules, a factor that contributes significantly to the architecture of the active site [3–6]. The subset that comprises non-haem iron(II)- and 2OG-dependent dioxygenases (i2OGdd) is characterized by the reaction chemistry, broad spectrum of substrates and is present in all kingdoms of life. The enzymes are characterized by a triad of Histidine, Aspartic-/Glutamic-acid and Histidine residues which co-ordinate ferrous iron. 2OG and the substrate. Enzymes are classified on the basis of the substrate(s) transformed and dominant reaction chemistry. Each cluster is a HMM-profile of at least two members and is derived from enzymes with available empirical data (structure, kinetic, mRNA expression). ALKB Alk-B like demethylase, ARG1 Arginine hydroxylase; ASPA Aspartyl/Asparaginyl hydroxylase, CHLO Chlorinating enzyme, CLAS Clavamine synthase; COLY Collagen lysyl dioxygenase, CP3H Collagen prolyl 3-hydroxylase, CP4H Collagen prolyl 4-hydroxylase, CYCL Cyclization and ring closure, DACS Deacetoxycephalosporin-C synthase, DSAT Desaturases, ECTO Ectoine hydroxylase, FLAV 2S-Flavones, GBBH γ-butyrobetaine hydroxylase, GIAC Gibberellic acid modification, HP4H Hypoxia prolyl 4-hydroxylase, HYOS Hyoscyamine hydroxylase, MGUT Nucleotide/nucleoside hydroxylase, OGD/E Eukaryotic initiation factor 2a, PHYT Phytanoyl-CoA hydroxylase, PTBL 1-Deoxypentalenic acid 11β-hydroxylase, SULF Sulfate cleaving, TFDA 2,4-Diphenoxycetic acid metabolizing, THYD Thymidine dioxygenase, THYE Thymine dioxygenase, XANT Xanthine hydroxylase

sphere with iron(II) [3–6]. The other face is formed by three displaceable water molecules, a factor that contributes significantly to the architecture of the active site [3–6]. The subset that comprises non-haem iron(II)- and 2OG-dependent dioxygenases (i2OGdd) is characterized by a triad of Histidine, Aspartic-/Glutamic-acid and Histidine residues which co-ordinate ferrous iron. 2OG and the substrate. Enzymes are classified on the basis of the substrate(s) transformed and dominant reaction chemistry. Each cluster is a HMM-profile of at least two members and is derived from enzymes with available empirical data (structure, kinetic, mRNA expression). ALKB Alk-B like demethylase, ARG1 Arginine hydroxylase; ASPA Aspartyl/Asparaginyl hydroxylase, CHLO Chlorinating enzyme, CLAS Clavamine synthase; COLY Collagen lysyl dioxygenase, CP3H Collagen prolyl 3-hydroxylase, CP4H Collagen prolyl 4-hydroxylase, CYCL Cyclization and ring closure, DACS Deacetoxycephalosporin-C synthase, DSAT Desaturases, ECTO Ectoine hydroxylase, FLAV 2S-Flavones, GBBH γ-butyrobetaine hydroxylase, GIAC Gibberellic acid modification, HP4H Hypoxia prolyl 4-hydroxylase, HYOS Hyoscyamine hydroxylase, MGUT Nucleotide/nucleoside hydroxylase, OGD/E Eukaryotic initiation factor 2a, PHYT Phytanoyl-CoA hydroxylase, PTBL 1-Deoxypentalenic acid 11β-hydroxylase, SULF Sulfate cleaving, TFDA 2,4-Diphenoxycetic acid metabolizing, THYD Thymidine dioxygenase, THYE Thymine dioxygenase, XANT Xanthine hydroxylase

The work presented revises, updates and integrates the functionality of two servers, i.e., Fe(2)OG ≡ \{H2OGpred, DB2OG\} [18, 19]. Fe(2)OG can be used by researchers as a single-point web resource to screen protein sequence(s) for potential i2OGdd-activity and shortlist putative i2OGdd members from the available pre-compiled sequence repository. The latter is searchable on the basis of taxonomy, cellular compartment and HMM-profiles of the sequences. A novel feature of Fe(2)OG is the inclusion of clinically relevant non-haem iron(II)- and 2OG-dependent dioxygenases. This includes links and preliminary analyses to several
Main text

Rationale for incorporating empirical data into a profile-based search application

Non-haem iron(II)- and 2OG-dependent-dioxygenases are characterized by variable reaction chemistry and a broad spectrum of substrates. The reverse mapping of substrate descriptors to the active site of known enzymes is well documented and can be utilized to repurpose pharmacological agents. Several theoretically sound statistical tools such as multi-class support vector machines (SVMs), artificial neural networks (ANNs), and hidden markov models (HMMs) have been utilized to garner insights into the active site geometry of an enzyme in the presence of a pharmacophore [18–22]. Although HMMs, as a predictive modality are non-committal, this can be rectified by mathematical filters. The transformed output can then be utilized by clustering algorithms and ANNs to generate unambiguous predictors [21, 23, 24]. In fact, a rigorously derived integrated HMM-ANN algorithm has been presented and used to characterize sequences which are few and closely related such as those from an enzyme family or sub family [23, 24].

Mathematical basis for the algorithms deployed by Fe(2)OG

Whilst, a detailed description of the computational pipeline deployed and its relevance has already been published, the mathematical basis for these has not been addressed [18, 19]. Briefly, HMM-profiles of catalytically relevant clusters and laboratory validated enzymes of the i2OGdd-superfamily \((a_i \in A \subseteq \mathcal{H})\) are utilized to score regions of an amino acid sequence. The empirical data that is considered is the presence of one or more 3D-structures, kinetic and mutagenesis data and mRNA expression levels [18]. A suitable mathematical representation is as under:

\[
A = \{a_1, a_2, \ldots, a_i | #a_i \geq 2\} = \bigcup a_i
\]

**Theorem:** A unique set of HMM profiles \((A, B \subseteq \mathcal{H})\) can exist iff there is at least one unique sub-profile.

\[
A \neq B \iff \#(a_i \cap b_i) \leq \min(#a_i, #b_i) \quad a_i \in A \subseteq \mathcal{H}, b_i \in B \subseteq \mathcal{H}, \{#a_i, #b_i\} \geq 2
\]

**Proof:**

| Case1  | if \(\cap_{i=1} a_i = \emptyset, \cap_{i=1} b_i = \emptyset\) |
|--------|--------------------------------------------------|
| Since\(A, B \subseteq \mathcal{H}\) | \(A \cap B \neq \emptyset\) |
| Rewriting, Expanding, Let, | \((\cup_{i=1} a_i) \cap (\cup_{i=1} b_i) \neq \emptyset\) |
| Clearly, Generalizing, Case2 Let, | \(\exists (a_i \cap b_i), \#(a_i \cap b_i) \leq \min(#a_i, #b_i)\) (a) |
| if \(\cap_{i=1} a_i \neq \emptyset, \cap_{i=1} b_i \neq \emptyset\) | \(\emptyset \equiv= \cup a_i \cap b_i\) |
| if\(\cap_{i=1} a_i = a_i \in A | a_i \geq 2, \cap_{i=1} b_i = b_i \in B | b_i \geq 2\) | \(#(a_i \cap b_i) \leq \min(#a_i, #b_i)\) (b) |
| From(a), (b) | \(A \neq B\) |

Conversely, \(A \neq B\), \(A \cap B \neq \emptyset\)

Since\(A, B \subseteq \mathcal{H}\), \(A \cap B = a_i \cup b_i\)

\(\Rightarrow \exists (a_i \cap b_i) \in A \cap B | #(a_i \cap b_i) \leq \min(#a_i, #b_i)\)

\(a_i \neq b_i\)
Fe(2)OG, then, is an implementation of a particular instance of the combined HMM of sequences and available structures (\( A = a_i \mid 1 \leq i \leq 28, 2 \leq \#a_i \leq 4 \)) [19]; URL-http://janelia.org. The lower limit of number of the sequences in each profile (\( \text{min}(\#a_i) \)) Eq. (1) is implied by definition. The upper limit, however, is estimated as a proportion of the total number of sequences,

\[
\max(\#a_i) = \frac{\#A}{\#H}
\]

Description and utilization of Fe(2)OG

**Fe(2)OG**, a predictor of the catalytic spectrum of an unknown or single function enzyme

The algorithm and code that Fe(2)OG utilizes to predict the dominant profile, in a user-defined sequence(s), has been described in detail [18]. Briefly, i2OGdd enzymes (\( n > 220 \)) with available empirical data (structure, kinetic, mRNA expression) are clustered on the basis of the substrates catalyzed and/or the reaction chemistry (Fig. 1) [18]. The enzymes present in each
Table 1  Comparative analysis and biomedical relevance of non-haem iron(II)- and 2OG-dependent dioxygenases

| A | Nature and composition | Mode of prediction and impact | Putative i2OGdd sequences | Display and accessibility | "Unclassified" sequences | "All sequences" search criteria | Clinically relevant members |
|---|------------------------|-----------------------------|---------------------------|--------------------------|--------------------------|--------------------------------|---------------------------|
| | Integrated webserver   | Generic, improved and extendible | Recent and relevant | Matrix format (compartments, taxonomy) | Yes; Amenable to further investigation | Omitted; extended profiles of sequences with known function | Yes |
| | | | | | | | |
| B Enzymes (EC1.14.11.x) | Physiological role; Disease biology | Profile(s); References |
| 1 | Phytanyol-CoA hydroxylase ($x = 18$) | Phytanic acid hydroxylation; Hypoxia-induced Proline hydroxylation | Refsum's disease | PHYT, [18, 27] |
| 2 | (HIF) Prolyl hydroxylases ($x = 2, 29$) | Hypoxia-induced Aspartic acid/ Asparagine hydroxylation | Tumor suppression, ubiquitin-mediated proteosomal degradation of HIF via the Von Hippel-Lindau complex | HP3H, HP4H, [13, 14, 18, 28] |
| 3 | (HIF) Asparaginyl hydroxylases ($x = 16, 30$) | Hypoxia-induced Aspartic acid/ Asparagine hydroxylation | Traboulsi's syndrome | ASPA, [18, 29, 30] |
| 4 | Collagen Prolyl hydroxylases ($x = 2, 28$) | Assembly of mature collagen | Connective tissue disorders, promoter of metastasis | CP3H, CP4H, OGFD; [18, 31, 32] |
| 5 | Procollagen-lysine 5-dioxygenase ($x = 4$) | | Kyphoscoliotic Ehlers–Danlos syndrome | COLY; [18, 33, 34] |
| 6 | JMJD6 (Jumonji-domain containing protein) ($x = 66, 67$) | Arginine demethylation, Lysine hydroxylation | Promoters of metastasis, Developmental disorders, | HILY, [18, 35, 36] |
| | Histone H3 trimethyl L lysine (4/9) demethylase | Lysine demethylation | | |
| | Histone H3 dimethyl L lysine 36 demethylase ($x = 27$) | | | |
| 7 | DNA demethylases ALKBH-1, 2, 3, 4, 5, 8 ($x = 34, 31, 32, 33$) | DNA/mRNA-repair after alkylation | Heightened propensity for malignant transformation, disorders of growth and development | ALKB, [18, 37, 38] |

'functional'-group, ($2 \leq nA_i \leq 4$) Eqs. (1) and (2) are then aligned and assigned a HMM-profile (Figs. 1 and 2) [18]. A database of these HMM-profiles is used to probe the catalytic spectrum of a user-defined sequence as per the stringency specified. Unlike H2OGpred, Fe(2)OG, compares a query sequence(s) with all, rather than isolated HMM-profiles (Fig. 2) [18]. The rationale for this alteration is that since the catalytic profile of an unknown sequence(s) is debatable, a generic analysis rather than a specific one is a better indicator of i2OGdd-like activity. Furthermore, sequences with known function can also be investigated for other reaction chemistries. Clearly, in both cases the analysis with individual profiles is superfluous and may be omitted (Table 1A). The tabulated list of relevant cognate substrates, for each profile is also available and may be used as a reference (Figs. 1 and 2). In addition, to the overt directives of use, users can also sample the functionality of Fe(2)OG by clicking the "Examples" button (StepP1) (Fig. 2). This loads bonafide i2OGdd sequences into the text area which can be analyzed in accordance with the steps that are outlined subsequently. These include choice of threshold parameter (Evalue, Bitscore) and assignment of a suitable numerical value (StepsP2, P3) (Fig. 2). The output comprises a tabular summary of suitably matched profiles with detailed statistics and exhaustive pair-wise alignments of all supra-threshold matches (Fig. 2). Since, Fe(2)OG has dual functionality, the user can submit this independently (StepsP1 – P3 → Submit) (Fig. 2).

ii) Fe(2)OG, a repository of i2OGdd-like sequences

The second component of Fe(2)OG is a flat-file database. This comprises a pre-compiled and updated list of i2OGdd-like sequences ($n_{AB} = 4496$) (Fig. 2). This is accomplished by constructing a generic-HMM after combining representative ($n \sim 80$) i2OGdd enzymes from each ‘functional’-group. This is then used to query UniprotKB for probable matches ($n_{AB}$) [19]. The downloaded sequences are analyzed and assigned a dominant cellular compartment ($n_A$) [19]. Sequences, which are not amenable to these preliminary investigations are annotated as such ($n_B$). Users can download updated lists of these sequences ($n_A = 3429, n_B = 1067$) (Fig. 2). This is facilitated by arranging the sequences as a matrix of compartments ($p$) and taxonomy.
$q \left( AB = \{ y_{pq} \in \{ ab_{pq} \} : p = 10, q = 7, r \in N \} \right)$. Fe(2)OG, also uses the logical operators (\{AND, OR\}) to formulate an advanced HMM profile-based query to partition the sequences (StepS1; Fig. 2) [19]. Another modification introduced in Fe(2)OG is the omission of the “All sequences”-option (StepS1) (Fig. 2). The rationale for this amendment, is that users may require sequences specific to one or more HMM-profiles (Figs. 1 and 2). Since, each profile is based on a specific reaction chemistry, users will also possess, a priori, a definitive list of probable ligands to characterize the kinetics of their search result with (Fig. 1, Table 1A). Furthermore, the entire database (nA) is accessible with the “OR” and “Include these profile(s)”, if the user so chooses (StepsS1,S2) (Fig. 2). The other fraction could not be further classified and is presented only in terms of their respective taxonomies (nB = 1067). Here, too, the user can submit this independently (StepsS1,S2 \rightarrow Submit) (Fig. 2).

**Comparative analysis and biomedical relevance of Fe(2)OG**

Despite the similarity in algorithms and general usage, Fe(2)OG, offers several new and upgraded features (Table 1). These include links to i2OGdd members which are uncharacterized and clinically relevant, whilst offering researchers a tool to extend the catalytic profiles of known enzymes. Additionally, the list of sequences with putative i2OGdd function is updated and non-redundant. The i2OGdd are amongst the largest group of non-haem dioxygenases and can arguably compete in importance with the more established cytochrome P450 (CYP) superfamily of haem monooxygenases (Fig. 1). The differential activity of i2OGdd members in response to fluctuating concentrations of oxygen and iron also suggest a system-level function in sensing and thence regulating the uptake, utilization and release of these micronutrients [25, 26]. In fact, clinical data is available for several i2OGdd enzymes. This includes phytanoyl-CoA hydroxylase, hypoxia-inducible Proline hydroxylases, collagen modifiers (Proline- and Lysine-hydroxylases) and DNA/mRNA-demethylases (Table 1B) [27–38]. The analysis by Fe(2)OG results in a small subset ($\approx 24\%$, $n = 17$) of enzymes and are grouped into mitochondrial, cytosolic and extracellular fractions (Additional file 1: Text S1a). However, a larger proportion ($\approx 76\%$, $n = 53$) remains unclassified and merits a deeper investigation (Additional file 1: Text S1b).

**Limitations**

Fe(2)OG, is an online web resource that is dedicated to expanding the knowledge base of non-haem iron(II)- and 2OG-dependent-dioxygenase superfamily of enzymes amongst scientists and clinicians. Fe(2)OG, can predict whether an unknown protein sequence(s) possesses i2OGdd-activity. It also provides preliminary analyses (taxonomy, cellular compartment) and an analytic tool (sequence-based, logical) to shortlist enzyme candidates from a pre-compiled list of sequences. Since, newer sequences are constantly becoming available, Fe(2)OG will require constant updates to its core of HMM-profiles and the raw sequences that are queried for putative function, thereof, to remain relevant to the biomedical community. However, since this information is dependent on available empirical data, an annual update might suffice. Fe(2)OG, is also not exhaustive and lacks structural-models and simulation data for its members. These short comings will be addressed in future studies.

**Supplementary Information**

The online version contains supplementary material available at https://doi.org/10.1186/s13104-021-05477-z.

**Additonal file 1: Text S1.** a HMM profiles of analyzed human i2OGdd-like sequences ($n = 17$); b Uniprot IDs of unprofiled human i2OGdd-like sequences ($n = 53$)

**Abbreviations**

Fe: Iron; 2OG: 2-Oxoglutarate; HMM: Hidden Markov Model; i2OGdd: Non-haem Iron(II)- and 2-oxoglutarate-dependent dioxygenases.

**Acknowledgements**

Not Applicable.

**Authors' contributions**

SK outlined and designed the study, designed and conceptualized the algorithm(s) and formulae for prediction, wrote the mathematical proofs, manually collated all the sequences, and their references, carried out the computational analysis, wrote all the code and the manuscript. All authors read and approved the final manuscript.

**Funding**

This work is funded by an early career intramural grant awarded to SK (Code A-766) by the All India Institute of Medical Sciences (AIIMS, New Delhi, INDIA).

**Availability of data and materials**

Data is available as supporting material with the manuscript.

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Yes.

**Competing interests**

The authors declare that they have no competing interests.

**Received:** 11 July 2020  **Accepted:** 3 February 2021  **Published online:** 01 March 2021

**References**

1. Koehntop KD, Emerson JP, Que L Jr. The 2-His-1-carboxylate facial triad: a versatile platform for dioxygen activation by mononuclear non-heme iron(II) enzymes. J Biol Inorg Chem. 2005;10(2):87–93.
22. Kundu S. Insights into the mechanism(s) of digestion of crystalline cellulose by plant class C. GH9 endoglucanases. J Mol. Model. 2019;25(8):240.

23. Kundu S. Mathematical basis of predicting dominant function in protein sequences by a generic HMM-ANN algorithm. Acta Biotheor. 2018;66(2):135–48.

24. Kundu S. Mathematical basis of improved protein subfamily classification by a HMM-based sequence filter. Math Biosci. 2017;293:75–80.

25. Wilson JW, Shakir D, Batie M, Frost M, Roche S. Oxygen-sensing mechanisms in cells. FEBS J. 2020. https://doi.org/10.1111/febs.15374.

26. Kundu S. Co-operative intermolecular kinetics of 2-oxoglutarate dependent dioxygenases may be essential for system-level regulation of plant cell physiology. Front Plant Sci. 2015;6:489.

27. Mukherji M, Chien W, Kershaw NJ, Clifton U, Schofield CJ, Wierzbicki AS, et al. Structure-function analysis of phytanoyl-CoA 2-hydroxylase mutations causing Refsum’s disease. Hum Mol Genet. 2001;10(18):1971–82.

28. Hirsila M, Koivunen P, Gunzler V, Kivirikko KI, Myllyharju J. Characterization of the human prolyl 4-hydroxylases that modify the hypoxia-inducible factor. J Biol Chem. 2003;278(33):30772–80.

29. Patel N, Khan AO, Mansour A, Mohamed JY, Al-Assiri A, Haddad R, et al. Mutations in ASPH cause facial dysmorphism, lens dislocation, anterior-segment abnormalities, and spontaneous filtering blebs, or Traboulsi syndrome. Am J Hum Genet. 2014;94(5):755–9.

30. Pfeffer I, Brewitz L, Krojer T, Jensen SA, Kochan GT, Kershaw NJ, et al. Structural and mutational analysis of Escherichia coli AlkB provides insight into substrate specificity and DNA damage searching. PLoS ONE. 2010;5(1):e8680.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.